

Object Recognition Datasets and Challenges: A Review

Aria Salari^a, Abtin Djavadifar^a, Xiang Rui Liu^a, Homayoun Najjaran^{a,*}

^a*School of Engineering, University of British Columbia, 1137 Alumni Ave, Kelowna, V1V 1V7, BC, Canada*

Abstract

Object recognition is among the fundamental tasks in the computer vision applications, paving the path for all other image understanding operations. In every stage of progress in object recognition research, efforts have been made to collect and annotate new datasets to match the capacity of the state-of-the-art algorithms. In recent years, the importance of the size and quality of datasets has been intensified as the utility of the emerging deep network techniques heavily relies on training data. Furthermore, datasets lay a fair benchmarking means for competitions and have proved instrumental to the advancements of object recognition research by providing quantifiable benchmarks for the developed models. Taking a closer look at the characteristics of commonly-used public datasets seems to be an important first step for data-driven and machine learning researchers. In this survey, we provide a detailed analysis of datasets in the highly investigated object recognition areas. More than 160 datasets have been scrutinized through statistics and descriptions. Additionally, we present an overview of the prominent object recognition benchmarks and competitions, along with a description of the metrics widely adopted for evaluation purposes in the computer vision community. All introduced datasets and challenges can be found online at github.com/AbtinDjavadifar/ORDC.

Keywords: Computer vision, Object recognition, Deep learning

1. 1 Introduction

Object recognition is one of the fundamental computer vision tasks that pertains to identifying objects of different classes withing digital visual representations such as images or 3D point clouds. As one of the most important tasks in computer vision, it has found practicality in numerous applications, from autonomous driving and face recognition to human pose estimation and remote sensing. Throughout the lifetime of computer vision research, datasets have played a critical role in the improvement of the object recognition algorithms. Not only do datasets provide a means to compare and contrast existing algorithms, but they also help push the boundaries for meaningful visual understanding as they evolve in the scale and diversity of covered scenarios. With the emergence of deep learning techniques [128], the need for large and precisely annotated datasets has further been accentuated. In recent years, object recognition tasks have been transformed in terms of accuracy and complexity, moving from classification of iconic view single-object [176] images to instance-level segmentation of highly cluttered environments [55, 152]. A crucial part of this progress is due to the availability of large-scale and finely annotated datasets.

1.1. Comparison with Relevant Reviews

Following the breakthroughs in object recognition in recent years, many notable surveys have been published discussing progressions in algorithm development [155, 280, 271], and their practicality in various real-world applications [227, 78, 134, 146, 229]. While a number of these surveys provide valuable discussions on relevant datasets [81, 280], there is still a lack of dedicated surveys providing an in-depth analysis on

*Corresponding author; Tel.: +1-250-826-1621;
Email address: homayoun.najjaran@ubc.ca (Homayoun Najjaran)

the recent progressions and trends in object recognition from the dataset development standpoint. As the striking success of deep learning-based object recognition algorithms in recent years is largely dependent on the availability of large and carefully annotated datasets, an independent survey on relevant datasets can provide a more comprehensive understanding of the current challenges in object recognition research as a whole and serve as a roadmap for future research directions.

1.2. Scope

This paper aims to analyze the progression of prominent object recognition datasets in the past two decades, with a focus on the more evolved entries in recent years. Along with the datasets, related competitions and common evaluation metrics used for the assessment of recognition algorithms have also been discussed. Due to the wide breadth of object recognition datasets and limited space, we restrict our analysis to datasets containing still 2D images in the visible spectrum. Further, we narrow our coverage to annotated datasets since the challenging tasks in the field are mostly tackled with supervised learning approaches. We understand the importance of other data modalities, and also computer vision tasks that use object recognition as a component, but the discussion of any material other than the discussed limits of this paper would be beyond the scope of any reasonable length publication. Finally, where possible, we tried to gauge importance and popularity by referring to highly cited works presented in the the most accredited publication venues in the field such as CVPR, ICCV, etc. We acknowledge there are certainly valuable research efforts beyond the publications discussed in this paper that we had to ignore due to limited space.

1.3. Contributions of the Paper

A comprehensive survey evaluating the utility of object recognition datasets in a quantitative manner will be an important contribution to the field. By providing a technical overview of the historical developments and contrasting the state-of-the-art in critical frames such as data bias, collection sources, annotation quality, readers will be able to form a hierarchical understanding of the role of data in object recognition. Further, by providing concise tabular descriptions of the most prominent datasets in popular applications, this paper can serve as a starting point for practitioners and researchers in the field looking to narrow their search for appropriate datasets. Finally, as object recognition algorithms mature, the existing datasets become saturated, thus making it necessary to develop more challenging datasets. To this end, we analyze the current trends to provide researchers with insight regarding future dataset collection directions.

2. Background

In this section, a history of the prominent datasets evolving through time, and an introduction to milestones in object recognition techniques before and after the emergence of deep learning models are provided. The most eminent evaluation metrics are also discussed.

2.1. An Overview of Object Recognition Tasks

Before introducing various object recognition algorithms, it is noteworthy to discern different but similar computer vision tasks like image classification, object detection, object localization, instance segmentation, and semantic segmentation, to avoid any confusion that may happen later. An image classification method assigns a class label to a whole image, whilst object localization draws a bounding box around each object present in the image. Object detection is a combination of these two tasks and assigns a class label to each object of interest after drawing a bounding box around Object recognition is a general term used for referring to all of these tasks together [206]. Figure 1 shows how these concepts are related to each other.

By applying object detection models, we will only be able to build a bounding box for every object present in the image. However, it will not tell anything about the object’s shape as the bounding boxes are either rectangular or square in shape. Image segmentation models, on the other hand, will create a mask in pixel level for every object that appeared in the image. This method provides us with a more comprehensive understanding of the object(s) present in the image. The image segmentation task itself can be performed in two ways: 1) semantic segmentation which labels each pixel of the image with class of the object surrounded inside the pixel (class-aware labeling), 2) instance segmentation which identifies object boundaries in pixel level and distinguishes between separate objects from the same class (instance-aware labeling).

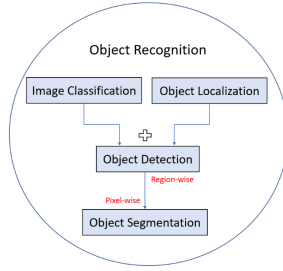


Figure 1: Overview of object recognition tasks in computer vision research

2.2. Historical Object Recognition Milestones

In the early developments of data-driven learning algorithms in the 90’s, object recognition algorithms mainly relied on sophisticated hand-crafted features designed for specific applications. As a result, researchers tried to create ad-hoc datasets for their specific research problem, and collective efforts for data collection and annotation were not as common as today. Early datasets usually focused on one particular application, face detection [183, 225, 113], face recognition [189, 20], pedestrian detection [184], and handwriting detection [24, 140] being among popular ones. Images typically were of low resolution and lacked high level annotations such as segmentation masks, and the objects in question were spatially well-defined. Cropping was also a widely adopted technique to present canonical views of the target object. Here, a canonical or iconic view refers to a clear and distinctive depiction of an object category in an image. Classification is the most frequent type of annotation in earlier datasets.

MNIST [139], sometimes referred to as the “Hello World” dataset of machine vision, is an example of the early image classification datasets, where each sample contains one hand-written digit in an uncluttered background. COIL-20 [176] and COIL-100 [175] are two other datasets of common objects with different orientations in a black background. FERET [189] is the cornerstone of face recognition datasets by offering a large collection of face galleries and a testing framework tuned to the accuracy levels of the then state-of-the-art algorithms [189, 114]. With the introduction of milestone feature descriptors such as SIFT [161], and Viola Jones [240, 239], and later on HOG [58], SURF [13], and DPM [77], datasets grew in size and object class coverage. Caltech-101 [75] and its successor Caltech-256 [92] made a striking turn by including objects in their natural environments, but these datasets were still limited to one class per image. CIFAR-10 and CIFAR-100 [127] are 10 and 100-class subsets of the Tiny Images dataset [233] in 32×32 pixel resolution. Numerous annotation errors occur in these datasets because the automatically generated annotations were not manually verified. In the early 2000’s, several attempts were also made to create datasets with higher quality of annotations. The first prominent dataset with segmentation mask annotation was BSDS [165], where human annotators manually segmented each image. An error measure was also introduced to evaluate the performance of segmentation algorithms against ground truth human annotations. However, the size of the dataset is by orders of magnitude smaller than the modern-day segmentation datasets, such as Microsoft COCO [152], and the evaluation criteria predicts zero error in cases where there is a proper subset relationship between the predicted and ground truth segments. LabelMe [207] leveraged crowdsourcing to acquire polygon annotation for objects in natural contexts but lacked uniform annotations because of arbitrary captioning and inaccurate polygons of annotators [206]. However, the online LabelMe annotation tool was extensively used to annotate other datasets, e.g., MIT-CSAIL [234], SUN [256], and ADE20K [274]. PASCAL VOC [69] was a stepping stone among the datasets of this era as it was adopted to hold annual object recognition competitions to benchmark the state-of-the-art algorithms. Up until 2012, algorithms utilizing local descriptors and hand tuned feature vectors dominated these competitions [68].

The current enthusiasm for deep learning-based methods goes back to 2012, wherein the winning entry of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) was a Deep Convolutional Neural Network (DCNN), i.e., AlexNet [128]. The use of CNNs for image classification dates back to the 1980s. The early works focused on the identification of hand-written digits, specifically as related to automated zip code detection [138, 137]. Research on the method continued through the late 1990s, but with little adoption

[140]. Lack of parallel compute power available at the time was the primary reason why CNNs were not used in large scale [195]. Advancement in GPU computation and increased availability of digitalized datasets led to a renewal of interest in 2006 and brought about technological advances such as the first application of maximum pooling for dimensionality reduction [40, 102, 103, 16, 194]. ILSVRC 2012 showed the full potential of deep algorithms can only be released when large-scale and accurately annotated datasets are available. Not only did DCNNs achieved staggering improvements in generic object recognition competitions, but they also opened doors for new applications where object recognition was never possible before. This led to a flurry of organized efforts to collect generic and application-specific datasets. With the resulting jump in recognition accuracy levels, datasets leaned more towards finer annotations, moving from classification of object classes to pixel-level annotations [152, 55]. Also, datasets were now being gathered in less controlled settings. Recently, there has been a focus to fully annotate images with a dense number of object instances in them to extract richer contextual information [94].

The evolution of milestone DCNNs emerging after AlexNet is discussed in the following. It needs to be mentioned that since a thorough analysis of these algorithms is beyond the scope of this paper, we only outline the development trajectory of these algorithms. For more information readers are encouraged to refer to relevant surveys [278, 156]. **R-CNN** [88] is a two-step Region-based Convolutional Neural Network. The R-CNN structure is based on two steps. First, selecting a feasible number of bounding-box object regions as candidates (also known as RoI) using a selective search approach. Second, extracting CNN features to perform classification while taking the features from each region independently. To overcome the expensive and slow training process of the R-CNN model, Girshick et al. enhanced the training process by joining three independent models together and training the new framework, called **Fast R-CNN** [87]. Although Fast R-CNN was noticeably faster during training and testing, the achieved improvement was not significant because of the high expenses of separately generating the region proposals by another model. **Faster R-CNN** [202] tried to speed up this process by integrating the region proposal algorithm into the CNN model by constructing a single, joined model made of fast R-CNN and RPN (Regional Proposal Network) having the same convolutional feature layers. **Single Shot Detector** (SSD) was introduced by C. Szegedy et al. in [158] in 2016 and could reach a new record in object detection task performance. Despite the RPN-based approaches like R-CNN series that generate region proposals and detect each proposal's object in two separate stages, SSD detects multiple objects present in the image in only one shot which lets it perform faster. Single Shot means that object classification and localization tasks are both done in a single forward pass of the CNN. SSD takes the bounding boxes' output space and discretizes them into a group of default boxes over various aspect ratios and then scales them for every feature map location. SSD proved its competitive accuracy against other techniques with an object proposal step by achieving better results on the MS COCO, PASCAL VOC, and ILSVRC datasets. Contrary to R-CNN family that locate the objects present in an image using regions by only looking at the areas of the images where are more probable to contain an object, **YOLO** [198] framework considers the whole image as its input and makes predictions about the coordinates of the bounding boxes and their probabilities. Three main advantages of YOLO are 1) its incredibly fast speed by processing images at 45 fps in real-time, 2) its ability to perform global reasoning and seeing the entire image while making predictions despite the region proposal-based and sliding window techniques 3) its capability to understand generalized object representations. After the introduction of YOLO [198] in 2016, various versions of YOLO have been developed trying to enhance its performance in different aspects. Fast YOLO is a version of YOLO using 9 convolutional layers instead of 24 which performs about 3 times faster than YOLO but has lower mAP (mean Average Precision) scores [210]. YOLOv2 is focused on reducing the significant number of localization errors and improving the low recall of original YOLO while maintaining classification accuracy [199]. YOLOv3 is the latest member of the YOLO family with some improvements compared to YOLOv2 such as a better feature extractor, a backbone with shortcut connections, and a more accurate object detector with feature map upsampling and concatenation [200]. Pixel-level segmentation needs more accurate alignment compared to bounding boxes. **Mask R-CNN** [96] was an effort to improve the RoI pooling layer to provide more precise mapping of the RoI to the regions of the original image. Mask-RCNN beat all the records in different parts of the COCO suite of challenges [152]. In 2019, a new framework called TensorMask [45] was introduced for extremely accurate instance segmentation tasks [45]. TensorMask uses a dense, sliding-window technique and novel

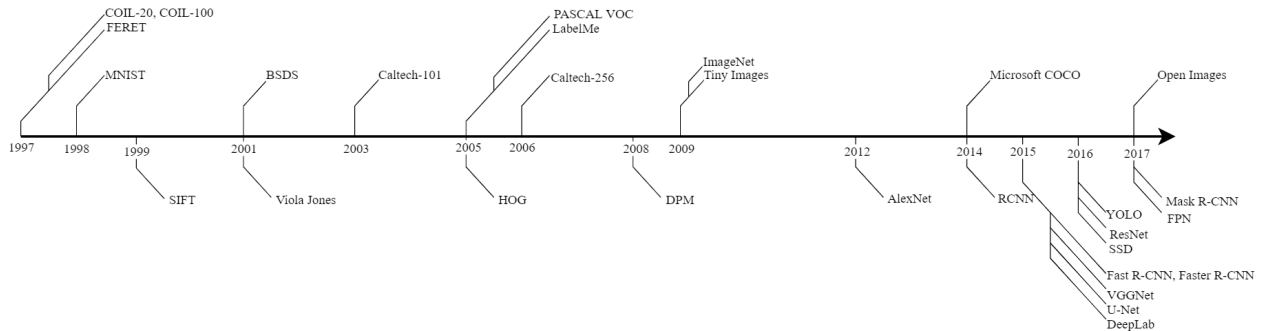


Figure 2: Milestones in object recognition algorithm and dataset development.(MNIST [139, 1], FERET [189], COIL-20 [176], [175], BSDS [165], Caltech-101 [75], Caltech-256 [92], LabelMe [207], Pascal VOC [69], Tiny Images [233], ImageNet [61], Microsoft COCO [152], Open Images [130], SIFT [161], Viola Jones[240, 239], HOG [58], AlexNet [128], R-CNN [88], Fast R-CNN [87], Faster-RCNN [202], VGGNet [38], U-Net [203], DeepLab [42], Mask-RCNN [96], FPN [151].

architectures and operators to capture the 4D geometric structure with rich and effective representations for dense images. **DeepLab** [42] is a revolutionary semantic segmentation model that uses atrous convolutions to simply upsample the output of the last convolutional layer and compute a pixel-wise loss to make dense predictions. DeepLab V1 added a number of advancements to the previous models by the employment of Fully Convolutional Networks (FCN) that led to overcoming two main challenges: 1) reduced feature resolution because of the multiple pooling and downsampling layers in DCNNs, 2) reduced localization accuracy because of DCNNs invariance. DeepLab V2 attempted to further enhance the performance of the DeepLab V1 by addressing the challenge of the existence of objects at multiple scales. It used a novel technique called Atrous Spatial Pyramid Pooling (ASPP), wherein multiple atrous convolutions with distinct sampling rates were applied to the input feature map and then outputs were combined [43]. DeepLab V3 was the last iteration of the DeepLab family that employed a novel encoder-decoder with atrous separable convolution which could obtain boundaries of sharper objects.[44]. **U-Net** [203] is an encoder-decoder based architecture first developed for biomedical image segmentation. One of the important challenges in the medical computer vision field is the limited number of datasets as many are not publicly available to protect patient confidentiality. Moreover, accurately annotating medical images requires trained personnel. U-Net employs data augmentation techniques to effectively use the available labeled samples which makes it a useful tool in cases without large annotated datasets. U-Net has also been shown to perform well on grayscale datasets.

2.3. Metrics

In order to dive into the common metrics used for the evaluation of computer vision algorithms, a few basic definitions need to be first established:

1. True Positive (TP): A correct prediction of a present condition
2. False Positive (FP): An incorrect prediction of a condition when it is not present.
3. True Negative (TN): A correct prediction of absence of a condition.
4. False Negative (FN): An incorrect prediction of absence of a condition when it is present.

Having defined the above terms, precision and recall could be defined as follows: Precision: Precision is a measure of the accuracy of the positive case predictions. A high precision represents a low level of wrong predictions (FP). It is defined as:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

Recall (Sensitivity, True Positive Rate): Recall quantifies the number of positive cases that have been correctly predicted by the algorithm. A high recall value shows a low level of undetected positive cases.

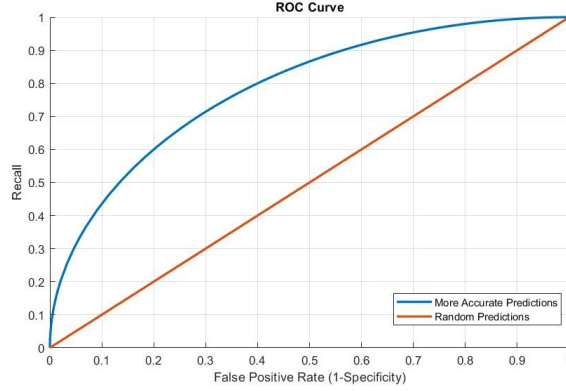


Figure 3: ROC Curve

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

Specificity (True Negative Rate): Specificity measures the share of actual negatives that have been correctly predicted as such.

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

Accuracy: Accuracy shows the ratio of correctly predicted instances to all the available instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

It needs to be noted that a similarity threshold between the annotated data and the predictions could be modified in algorithms to determine whether the prediction of a positive case is as accurate enough to be considered as a TP. Tuning the similarity threshold will in turn alter the precision and recall values. For instance, an increase in precision stemmed from a strict threshold comes at the expense of a reduction in the number of detected positive cases and hence recall. Receiver Operative Characteristic (ROC) Curve: Mostly used in the medical diagnostic and face verification problems, the ROC curve is obtained by plotting True Positive Rate (Recall) as a function of False Positive Rate (1-Specificity). Each point on the ROC curve corresponds to a set of sensitivity and specificity associated with a decision threshold. For a binary classification problem with equal number of data entries for each class, the curve will be a line of slope 1, given the predictions are entirely random. As the model starts to make better predictions, the curve is bent towards the upper left corner (Fig. 2). The area under the ROC curve is used as a performance metric. A higher area under the ROC curve indicates a higher detection rate of true positives (high sensitivity) while making less false positive predictions (high specificity), therefore specifying a better performance.

In the ROC curve, if the false positive rate in the horizontal axis is substituted with the number of false positive per image, the Free Operative Characteristic (FROC) is produced instead. F_1 Score: As previously discussed, these metrics serve different evaluation goals. The F_1 scores has been proposed to combine both precision and recall values into their harmonic mean:

$$F1 = 2 \frac{precision \times recall}{precision + recall} \quad (5)$$

Dice coefficient: First defined for discrete sets, the Dice coefficient is just reformulation of the F_1 score when applied to Boolean data with TP, FN, and FP definitions:

$$Dice = \frac{2TP}{2TP + FP + FN} \quad (6)$$

For a given detection/segmentation task, the numerator represents twice the area of the ground truth and prediction overlaps, and the denominator shows the total pixel area of the ground truth and the prediction. Intersection over Union (IoU, Jaccard Index): IoU is used for thresholding the TPs in objection detection and segmentation tasks. It is defined as the intersection of the ground truth segmentation/bounding box and the predicted segmentation/bounding box over the union of those the ground truth and predictions:

$$IoU = \frac{AreaofOverlap}{AreaofUnion} = \frac{TP}{TP + FN + FP} \quad (7)$$

Bad predictions (FP and FN) are more penalized in IoU than in the Dice coefficient, thereby the IoU score of an algorithm is always less than or equal to that of the Dice coefficient. However, in an overall sense these metrics are quite similar and positively correlated, meaning given a set of algorithms for a specific task, both metrics will rank their performance in the same manner. The IoU of different classes in an image entry could also be averaged to calculate mean Intersection over Union (mIoU). Average Precision (AP): Average Precision and its variations are probably the most popular metrics among computer vision challenges. AP essentially computes the precision of an algorithm over recalls ranging from 0 to 1. In order to calculate the AP of an algorithm for a specific class and over a dataset, the data entries need to be first ranked with respect to their confidence score. The confidence score of a class could be the outputted probability of a classifier, or the IoU score of a segmentation/object detection algorithm. The mean of the precision values for the top-k ranked predictions with k ranging from 1 to n (the number of positive ground truths) forms the average precision:

$$AP = \sum_{k=1}^n P(k) \Delta r(k) \quad (8)$$

Where $P(k)$ is the precision score of the top-k predictions, and $r(k)$ is the recall change between steps k and $k + 1$. It needs to be noted that the recall value at step k is defined as the number of TPs in the top k confident predictions over the total number of TPs in the dataset for the class in question. The above equation can be reformulated as:

$$AP = \frac{\sum_{k=1}^n P'(k)}{total\ number\ of\ TPs} \quad (9)$$

Where $P'(k)$ is defined in the same manner as $P(k)$ but returns zero is the k th prediction is not a true positive. To measure the performance over several classes instead of one, the AP score could be averaged to compute the mean Average Precision (mAP):

$$mAP = \frac{\sum_{c=1}^{C_{total}} AP(c)}{C_{total}} \quad (10)$$

Where C_{total} is the total number of classes. Panoptic Quality: The performance of panoptic segmentation tasks is measured using the proposed Panoptic Quality (PQ) metric. PQ incorporates both the detection and segmentation quality of algorithms and is formulated as:

$$PQ = \frac{\sum_{(p,q) \in TP} IoU(p,g)}{|TP|} \times \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \quad (11)$$

Where TP, FP, and FN correspond to the number of correct detections with $0.5 < IoU$ segmentation score, incorrect detections with $IoU > 0.5$, and unmatched ground truth segments (not detected or $0.5 > IoU$) respectively. Note that the first fraction is the average of IoUs for the matched segmentations and does not take the bad predictions into account whereas the F_1 score in the fraction to the right evaluates the detection quality with the $\frac{1}{2}|FP| + \frac{1}{2}|FN|$ term penalizing bad and missed predictions. Cancelling TP in the two fractions we can write the PQ score as follows:

$$PQ = \frac{\sum_{(p,q) \in TP} IoU(p,q)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \quad (12)$$

The PQ metric is computed for every class and then averaged over all classes.

3. Generic Object Recognition Datasets

The number of the publicly available labelled datasets has soared in recent years. Such datasets are usually annotated through crowdsourcing. Amazon Mechanical Turk is a popular crowdsourcing platform, used for many of the famous datasets [61, 126, 216]. For several of the datasets, there also exists a challenge to benchmark the performance of the state-of-the-art algorithms. In this section, a comprehensive overview of the datasets and the corresponding challenges focusing on the generic object recognition task have been provided. In Sec. 3.1, the four major large-scale object recognition datasets are discussed, with other generic datasets to follow in Sec. 3.2 and 3.3.

3.1. Major Large-scale Datasets

There are four major object recognition datasets widely adopted by researchers. The statistics of these datasets are provided in Table 1. Each dataset comes with a corresponding challenge held annually to evaluate and benchmark state-of-the-art algorithms. These challenges have played a critical role in the advancements of object recognition techniques in recent years, as the state-of-the-art algorithms emerge to outperform the winners from previous competitions. Figure 3 shows the rise in the performance of the winner algorithms in each of the four major challenges.

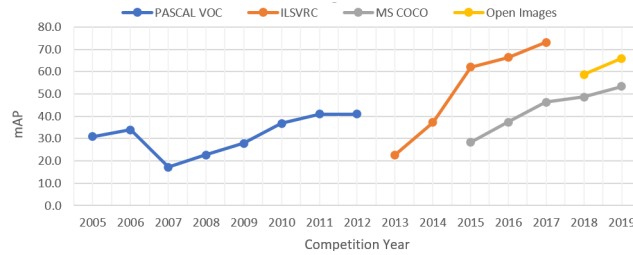


Figure 4: Accuracy improvement of winner algorithms in the object detection track of major challenges. The fall in accuracy levels in PASCAL VOC 2007 is due to the increase of object classes from 4 to 20.

Table 1: Dataset statistics for PASCAL VOC, ImageNet, MS COCO, and Open Images

Dataset	Number of Classes	Number of Images	Average Objects Per Image	First Introduced
PASCAL VOC	20	22,591	2.3	2005
ImageNet	21,841	14,197,122	3	2009
Microsoft COCO	91	328,000	7.7	2014
Open Images	600	9,178,275	8.1	2017

3.1.1. Overview of Datasets

Pascal VOC: The earliest of the four is the Pattern Analysis, Statistical Modelling and Computational Learning Visual Object Classes (PASCAL VOC) dataset [70]. It was first released in 2005 with four object classes, which were then expanded to 20 in the following years. The PASCAL VOC annual challenges (2005-2012) [68] laid the foundation of the standard object recognition evaluation metrics, widely adopted by future competitions. The dataset fell out of fashion after the release of a significantly larger dataset named ImageNet, but it is still sometimes used as a convenient benchmark of newly developed algorithms because of its relatively small size.

ImageNet: Following the footsteps of PASCAL VOC, the ImageNet dataset was introduced in 2010 [61]. The dataset contains 14 million annotated images in highly fine-grained categories. Samples are organized based on the hierarchical structure of WordNet [76] that is a lexical database that groups conceptually similar words into sets of cognitive synonyms (synsets). The database consists of 27 main classes such as fish, vehicles, and tools. Each of these classes are then divided into semantically arranged synset subcategories, resulting in over 21000 synsets. A subset of 1.2 million images of the dataset forms the baseline for the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [206], held between 2010 and 2017. When first introduced, the only task in the challenge was image classification, but single-object localization and object detection tasks were also added in the later competitions. Due to its large size and extensive class hierarchy, ImageNet is still used to pretrain complex models (e.g. VGG [38], ResNet [97]). However, ImageNet is one of the largest object recognition datasets to date although issues such as partial image annotations (only one annotated object per image) and lack of rich natural contextual information have instigated attempts to create more intricate datasets.

Microsoft COCO: Microsoft COCO [152] is an effort to address the problems of ImageNet by offering a relatively small (328,000 images) but a much richer dataset. In order to improve generalizability, images are annotated with segmentation mask for objects in their contextual backgrounds instead of canonical views. The main argument for the inclusion of contextual environments in the dataset is that detection of many object categories depends on whether they are depicted in their natural surroundings or not. Specifically, the algorithms trained on a particular dataset will be more generalizable if the objects are presented in their natural environment. By training a model on both PASCAL VOC and MS COCO, and then testing it on the other dataset, it was shown that although the algorithm trained on PASCAL VOC performed better, the performance drop of the MS COCO trained algorithm on the test dataset was less significant; this indicates its superior generalizability. There has also been a focus on balancing the number of instances per category, ensuring that a threshold of 5000 entries is surpassed for all object classes. Additionally, a supplementary COCO stuff dataset [32] has been released that provides segmentations for surrounding-level classes like grass, sky, and floor, which lack distinctive structure or parts. The downside of the COCO dataset is that the object categories are entry-level, and fine-grained categorization is ignored so that each category includes at least a sufficient number of samples.

Open Images: Open Images [130] is the most recent large-scale dataset that includes more than nine million labeled images, with bounding boxes for 600 object classes. Segmentation masks were also added to the 2019 release of the dataset. Moreover, a crowdsourced extension adds 478,000 labeled, but not segmented, images to the dataset. Another feature of the dataset is visual relationship annotations that capture meaningful interactions between pairs of objects, allowing for more holistic interpretations. Image classes include coarse-grained object classes, fine-grained object classes, events, scenes, materials and attributes (e.g., texture and color). Image class labels are divided into positive labels, instances that exist in the image, and negative labels, instances that do not appear in the image. By running classifiers on the dataset with and without the help of negative labels, it was shown that inclusion of negative labels considerably enhances classification accuracy.

3.1.2. Challenge Tasks

In this section we analyze the development of the popular object recognition tasks covered in the corresponding challenges of the datasets discussed in section 3.3.1. Table 2 lists the criteria, dataset attributes, and performance metrics of the four annual challenges. Each challenge consists of multiple tasks including classification, object detection and segmentation elaborated in the following.

Classification: This task is only included in PASCAL VOC and ILSVRC challenges. More recent challenges do not include classification due to its relative simplicity. In PASCAL VOC, two competitions were organized. Participants could either train their models on the annotated VOC training/validation dataset, or use their pre-trained models in a separate competition. In ILSVRC, algorithms were tested on a list of images, each having one ground-truth annotation. Since there might be several objects in a single image, it was possible that the algorithm could not distinguish the desired object for labeling. As a result, for each image, the algorithm was allowed to label up to five objects, and the evaluation would then be

Challenge	Tasks Covered	Classes	Images	Annotated Objects	Years active	Task Description	Evaluation Metric
PASCAL VOC	Image Classification	20	11,540	27,450	2005 - 2012	Absence/presence prediction of at least one instance of every class in each image	AP
	Detection	20	11,540	27,450	2005 - 2012	Bounding box prediction for every instance of the challenge classes present in images	AP with $IoU > 0.5$
	Segmentation	20	2913	6929	2007 - 2012	Semantic segmentation for the object classes	IoU
	Action Classification	10	4588	6278	2010 - 2012	Bounding box prediction or single points for persons performing an action and annotate with the corresponding action label	AP over action classes
	Person Layout Taster	3	609	850	2007 - 2012	Body part (hands, head, feet) detection with bounding boxes	AP calculated separately for parts with $IoU > 0.5$
ILSVRC	Image Classification	1000	1,331,167	1,331,167	2010 - 2014	Classification for one annotated class per image	Binary class error over the top 5 predictions per image
	Object Localization	1000	573,966	657,231	2011 - 2017	Bounding box prediction for only one object per image	Binary class and bounding box IoU error over the top 5 predictions
	Object Detection	200	476,688	534,309	2013 - 2017	Bounding box prediction for all instances per image	AP with flexible recall threshold varied proportional to bounding box size
	Object Detection from Video	30	5,314 (video snippets)	-	2015 - 2017	Continuous bounding box prediction throughout video sequences	AP with flexible recall threshold varied proportional to bounding box size
MS COCO	Detection	80	123,000+	500,000+	2015 - present	Instance Segmentation over object classes (things)	AP at IoU range of [0.5 : 0.05 : 0.95]
	Keypoints	17	123,000+	250,000+	2017 - present	Simultaneous object detection and keypoint localization	AP based on Object Keypoint Similarity (OKS)
	Stuff	91	123,000+	-	2017 - present	Pixelwise segmentation of background categories	mIoU
	Panoptic	171	123,000+	500,000+	2018 - present	Full segmentation of images (stuff and things)	Panoptic Quality
	DensePose	-	39,000	56,000	2019 - present	Human body segmentation and mapping all the pixels of the body to a template 3D model	AP based on Geodesic Point Similarity (GPS)
Open Images	Object Detection	500	1,743,042	12,421,955	2018 - present	Hierarchical-based bounding box detection	mAP
	Instance Segmentation	300	848,000	2,148,896	2018 - present	Instance segmentation over object classes; negative labels included to refine training	mAP at $IoU > 0.5$
	Visual Relationship Detection	57	1,743,042	380,000 relationship triplets	2018 - present	Labeling images with relationship triplets containing the interacting objects and the action class	A weighted sum of mAP and recall of number of relationships at $IoU > 0.5$

Table 2: Challenge description for PASCAL VOC, ILSVRC, MS COCO, and Open Images

made based on whether the annotated object was listed between the predictions for that particular image. The error of an algorithm was measured with the top-5 criterion, defined as the fraction of test examples for which the algorithm does not classify the desired object in its top five predicted labels. The error is given by:

$$PQ = \frac{1}{n} \sum_k \min_j d(l_j, g_k) \quad (13)$$

where $d(x, y) = 0$ if $x = y$ and 1 otherwise. The top-5 criterion has been also been adopted by other classification datasets for benchmarking proposes [12, 105].

Object detection: All the four competitions offer this track. However, bounding box annotations will probably be sidelined for segmentation masks. As of 2018, this task is not featured in the MS COCO challenges. All competitions used a variation of mAP with IoU thresholding used to assess the algorithms. In essence, the difference between the evaluation criteria of these competitions is the way the IoU threshold is defined. PASCAL VOC and Open Images use a threshold of 0.5 whereas in ILSVRC the threshold value increases proportional to the size of each object. Since all the datasets have images containing unannotated object classes, predictions for such classes are not assessed.

Segmentation: This task is aimed for pixel-wise object classification of test images. All the competitions except for ILSVRC offer this track. In PASCAL VOC, models were required to classify every pixel as belonging to one of the 20 specified classes or background. A 5-pixel wide margin around objects were labeled as void so as to identify the border between the objects as and the background. Difficult objects (far, small, or highly occluded), were also segmented with a void mask and hence excluded from the training and the test datasets. Open Images follows the same evaluation metrics as its detection task. segmentation masks are evaluated using mAP with mask-to-mask or box-to-box IoU greater than 0.5 (depending on whether the prediction is tested against the ground-truth segmentation or bounding box) over 300 challenge classes. MS COCO tackles the segmentation task more comprehensively. Along with the standard instance segmentation, two other segmentation tasks are presented: stuff segmentation and panoptic segmentation. The former is aimed towards pixel-wise segmentation of the background-level categories such as grass and wall, instances of which are hard to separate from one another. Stuff classes are divided into two general outdoor and indoor categories, each of which are narrowed down in two more levels. The panoptic segmentation task is aimed for achieving fully segmented images by combining semantic segmentation (pixel-wise class labels) and instance segmentation (masking object instances). The performance of panoptic segmentation tasks is measured using the Panoptic Quality metric.

Competition-specific tasks: In addition to the three mainstream tasks mentioned above, there are competition-specific tasks that focus on a particular recognition application, e.g. human body part recognition or action classification. PASCAL VOC's Person Layout Taster is a human detection task reported with a confidence score. Body parts were also required to be identified with bounding boxes; the requirements for prediction

to be considered as a true positive were expanded to correct detection of body parts (legs, hands, and head) with their corresponding bounding boxes, having an IOU score of over 50%.

Open Images offers the Visual Relationship Detection, where participants are required to identify visual relationship triplets in two scenarios. In the first scenario, for each relationship class, algorithms generate two bounding boxes for the interacting objects, their corresponding labels, and a label for the visual relationship. In the second scenario, the output is a bounding box encapsulating the two interacting objects, and three labels, indicating the relationship triplet. Finally, a weighted sum of the recall of the first, and the mAPs of both scenarios is computed to rank the performance of the algorithms.

MS COCO’s Keypoint Detection is a human pose estimation task. It involves localization of important body keypoints that well describe the person’s pose. Different types of keypoints, such as eyes and hips, are labeled to describe person’s pose in unspecified and uncontrolled conditions. The accuracy of the keypoint detection algorithms is measured by calculating the mAP over a range of similarity criteria called Object Keypoint Similarity (OKS) defined as:

$$OKS = \frac{\sum_i e^{\frac{-d_i^2}{2s^2k_i^2}} \delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \quad (14)$$

where d is the Euclidean distance between the ground truth and the predicted keypoint, s is the scale parameter which determines the sensitivity of this similarity criterion to the surface of the keypoint, and k is the keypoint constant, which is used to take into account the importance of the keypoint type. For instance, since eye location is more important than hip location, $k_{eye} < k_{hip}$. Finally, v_i is the visibility factor which is zero for obscured ground-truth keypoints and 1 otherwise. This means that the occluded keypoints will not be used for benchmarking the performance of predictions. The mAP in the range of $OKS = [0.5 : 0.05 : 0.95]$ is then calculated to determine the best method in this task.

3.2. Other Object Recognition Datasets

Besides the datasets discussed in Sec. 3.1 many other object recognition datasets also exist that are used to a lesser extent but may still be useful for research and development purposes.

3.2.1. Object Detection Datasets

As previously discussed, the most important object detection datasets in terms of impact and size (at the time of their release) are PASCAL VOC and Open Images. However, there are numerous other recognition datasets with bounding box annotations. Many of these datasets, such as Caltech Pedestrian, and KITTI are application-specific and are therefore discussed in Sec. 4. Since the rapid improvement of state-of-the-art recognition algorithms demand for holistic scene annotations, there has been a shift in focus towards segmentation mask annotations. However, bounding boxes are still considerably less time-consuming to annotate and also well-defined, being more robust to inconsistencies arising from subjective annotations of different human annotators working on the same dataset [63]. Inconsistencies exist even between two segmentation annotations of the same image by one annotator [274]. Consequently, bounding box annotations are still used in many datasets. In Table 3, the information of some generic object detection datasets is provided. We ignored non-commercial, e.g. Google’s internal JFT-300m [51], or sparsely annotated (e.g., YouTube Objects [190], YFCC100M [230]) datasets.

Dataset	Number of Images	Classes	Number of Bounding Boxes	Year
Caltech 101 [75]	9,144	102	9144	2003
MIT CSAIL [234]	2,500	21	2500	2004
Caltech 256 [92]	30,307	257	30,307	2006
Visual Genome [126]	108,000	76,340	4,102,818	2016
YouTube BB [197]	5.6 m	23	5.6 m	2017
Objects 365 [211]	638,000	365	10.1 m	2019

Table 3: Generic object detection datasets (except for those covered in Sec. 3.1).

3.2.2. Object Segmentation Datasets

Object segmentation datasets offer either instance-level or semantic segmentation masks. Stuff classes are either not annotated or labeled as background. A list of impactful generic object segmentation datasets is provided in Table 4. Since stuff could also be segmented, there has been an inclination towards creating datasets with fully segmented images (these datasets are discussed in Sec.3.3). However, object segmentation datasets are still popular for Video Object Segmentation (VOS). Pixel-wise annotation of foreground objects with temporal correspondence between frames is vital to applications such as action recognition, motion tracking, and interactive video editing. Most VOC datasets are small in size, usually containing only a dozen of video sequences [112, 74, 142, 29, 180]. DAVIS [30] and YouTube-VOC [258] are two recent entries established as the main benchmarks for VOC. Corresponding challenges are held annually for both of the datasets, where algorithms are evaluated based on IoU and F_1 scores. As of 2020, an instance segmentation challenge is also held for the LVIS [94] dataset with the same AP evaluation metric as MS COCO.

Dataset	Number of Images	Classes	Number of Objects	Year	Challenge	Description
SUN [256]	130,519	3819	313,884	2010	No	The main purpose of the dataset is scene recognition, however instance-level segmentation masks have also been provided
SBD [95]	10,000	20	20,000	2011	No	Object contours on the train/validation images of PASCAL VOC
Pascal Part [46]	11,540	191	27,450	2014	No	Object part segmentations for all the 20 class in the PASCAL VOC dataset
DAVIS [30]	150 (videos)	4	449	2016	Yes	A video object segmentation dataset and challenge focused on semi-supervised and unsupervised segmentation tasks
YouTube-VOS [260]	4,453	94	7,755	2018	Yes	videos object segmentation dataset collected of short (3s-6s) video snippets
LVIS [94]	164,000	1000	2 m	2019	Yes	Instance segmentation annotations for a long-tail of classes with few samples
LabelMe[207]	62,197	182	250,250	2005	No	Instance-level segmentation; some of the background classes have also been annotated

Table 4: Object segmentation datasets.

3.3. Object Recognition in Scene Understanding Datasets

A comprehensive understanding of an image requires the determination of scene characteristics in addition to mere object recognition. Although the identification of object categories may provide some clues about the context of an image (e.g., beds and drawers in an image possibly represent a bedroom), but such information alone can be misleading since many common objects are shared across different contexts. Additionally, the amorphous background properties of an image (sometimes referred to as stuff categories versus objects/things) are ignored in object recognition datasets, despite being crucial in providing deeper visual information such as geometric relationship of the objects or contextual reasoning (e.g. trees are more likely to appear on grass rather than street). As a result of these shortcomings, many scene-centric datasets have been proposed in the computer vision community.

Scene recognition datasets are directly tailored for scene type identification by including classification labels for each of the images, describing the corresponding scene of that particular image. Scene is defined as a place where a human can navigate [256, 273]. Classification labels do not allow for intra-class variations, which can be significant in the case scene types (e.g., the variations between different forests). Additionally, multiple categories might exist in an image, and limiting the number of annotations to only one classification label will result in an inevitable loss of information for some scenes. In a few scene recognition datasets, additional annotations have been added to address these issues. In a subset of SUN database [256], different sub-scenes are annotated to allow for the detection of multiple scenes in one image. Creators of OpenSurfaces [15] took a different approach by annotating scene labels for segmented interior surfaces in each scene. An attribute-based categorization is used in SUN Attributes [187] to take into account the inter-class scene variations. The most popular scene recognition datasets are presented in Table 5. A 365-class subset of Places2 [276] was used for the Places scene classification challenge with a top-5 evaluation criterion similar to ILSVRC.

Scene parsing datasets are another type of scene understanding datasets that go beyond the recognition of object classes by including pixel-wise segmentations for all the stuff and object categories in an image. Compared to object recognition datasets, these datasets require more extensive class categorizations and much richer annotations, thus often sacrificing the scale. The majority of scene parsing datasets focus on outdoor scenes [55, 91, 231, 250]. Indoor scene parsing is still a more challenging task compared to outdoor scene parsing due to the higher diversity of scene categories and smaller amount of training data [232]. The

Dataset	Number of Images	Classes	Additional Annotations	Year	Description
15-Scene [135]	4,485	15	-	2006	One of the earliest major scene classification datasets
MIT Indoor67 [191]	15,620	67	-	2009	Indoor scene classification in 5 main groups: Store, Home, Public Space, Leisure, and Working Place
SUN [256]	130,519	899	313,844 SM (Objects)	2010	Classification dataset of navigable scenes with additional object recognition annotations
SUN Attribute [187]	14,000	700	102 binary attributes per image	2012	Attribute-based representation of scenes for a subset of the original SUN database
Open Surfaces [15]	25,357	160	71,460 SM	2013	Segmented surfaces in interior scenes with texture and material information
Places2 [273]	10 m	476	-	2017	Classification of scenes bounded by spaces a human body would fit; comes with binary attributes

Table 5: Popular scene recognition datasets.

prominent scene parsing datasets are summarized in Table 6. Among the listed datasets, scene parsing challenges have been held for ADE20K, MS COCO Stuff, and SUN RGB-D, all using IoU as the evaluation metric.

Dataset	Number of Images	Stuff Classes	Object Classes	Year	Challenge	Highlights
MSRC 21 [212]	591	6	15	2006	No	One of the earliest semantic scene parsing datasets; images were later used in PASCAL VOC and Stanford Background
Stanford Background [91]	715	7	1	2009	No	Outdoor scene parsing dataset collected from LabelMe, MSRC, and PASCAL VOC geometric features also included
SiftFlow [153]	2688	18	15	2009	No	An early dataset on outdoor environment scene parsing labeled using LabelMe
Barcelona [231]	15,150	31	139	2010	No	A subset of the LabelMe dataset
NYU Depth V2 [213]	1,449	26	893	2012	No	Parsing of 464 cluttered indoor scenes; depth maps also included; semantic segmentation masks for objects
SUN+LM [232]	45,676	52	180	2013	No	A fully annotated subset of LabelMe and SUN datasets with both indoor and outdoor images
PASCAL Context [171]	10,103	152	388	2014	No	Pixel-wise semantic segmentation on the PASCAL VOC dataset; 520 new object and stuff categories were added to the original dataset
SUN RGB-D [216]	10,335	47	800	2015	Yes	Indoor scene parsing dataset and benchmark; 3D bounding boxes also provided
Cityscapes [55]	25,000	14	13	2016	No	Images captured from a vehicle driving in urban environments across 50 cities in different weather conditions in Europe; instance-level segmentation masks
ADE20K [274]	25,210	1,242	1,451	2017	Yes	Includes object part labels, and attributes; instance-level segmentation masks
Synscapes [250]	25,000	14	13	2018	No	Photo-realistic synthetic scene parsing of urban environments; annotation categories are the same as Cityscapes instance-level segmentation masks
MS COCO Stuff [32]	163,957	91	80	2018	Yes	Pixel-wise semantic segmentation for the entire MS COCO dataset

Table 6: Scene parsing datasets.

Apart from generic object recognition datasets, a number of scene understanding datasets include annotations only for a portion of objects that provide meaningful information about the context of the scene. These include salient and camouflaged object detection datasets. Salient object detection refers to detecting the most salient objects within images. Salient objects form the most visually distinctive regions of an image and are most likely to be the first objects to catch the attention of human observers. As a result, accurate detection of such objects can provide a quick and efficient understanding of the context of the scene. Salient object detection has been widely studied throughout the years [149, 148], and a number of popular datasets have emerged throughout the years to fuel research in the field, see Table 7. Camouflage object detection on the other hand, is aimed to go beyond human perception capabilities by identifying object categories that are visually indistinguishable or hard to distinguish from background elements. Camouflaged object detection is still an understudied area primarily due to the inherent annotation difficulty and subsequent lack of large-scale datasets geared towards the task [72, 266]. The very few predominant datasets in the field include CHAMELEON [71], CAMO [136], and COD10K [72].

4. Fine-grained Object Recognition Datasets

One of the issues associated with generic object recognition datasets is class imbalance. Since many of these datasets are created using online sources, namely Flickr, there is a tendency to have a greater number of instances for categories that are more common on the web, such as airplanes, and cars. Additionally, large-scale datasets may lack a fine-grained categorization, and even if they are fine-grained (e.g., ImageNet), the number of instances naturally drops as we move down a hierarchical class categorization. As a result, the performance of an algorithm will suffer if it is trained on a generic dataset that is not as dense in the context in which the algorithm is used as it is in other categories. This problem can be addressed by fine-tuning the

Dataset	Number of Images	Objects per Image	Annotation Type	Year	Highlights
MSRA-A [157]	20,840	1-2	BB	2007	One of the earliest large-scale salient object detection datasets. 5,000 and 10,000 image subsets of the dataset were further created with pixel-wise annotations
ASD [3]	1,000	1	SM	2009	A subset of MSRA-A dataset with pixel-wise annotations
THUR15K [49]	15,000	1	SM	2013	Images are crawled from the web using butterfly, coffee mug, dog jump, giraffe, and plane keywords
DUT-OMRON [262]	5,172	1-4	BB	2013	Saliency detection dataset in complex environments, more challenging compared to MSRA
PASCAL-S [150]	850	1-5	SM	2014	A subset of the PASCAL VOC dataset. Also includes human gaze annotations
ECSSD [261]	1,000	1-4	SM	2015	Focused on providing images with challenging natural backgrounds
HKU-IS [143]	4,447	1-4	SM	2015	Low contrast images often containing multiple disconnected salient objects that touch the image boundary
SOS [268]	14,000	0-4+	BB	2015	Each image is labelled as having 0,1,2,3, or 4+ salient objects.
DUTS [242]	15,572	1-4+	SM	2017	Popular large-scale dataset with images from the ImageNet and SUN datasets
XPIE [253]	10,000	1-4	SM	2017	Objects are selected as salient if they have high objectness, topological simplicity, and less than four similar candidates within the image
SOC [73]	6,000	0-4	SM	2018	Half of the dataset does not contain salient objects. Object category and challenging factor annotations are also provided.
COCO-CapSal [269]	6,724	1-4	SM	2019	Contains 80 salient object categories collected from the Microsoft COCO dataset. Also includes human gaze annotations

Table 7: Popular salient object detection datasets.

algorithm on datasets with richer annotations in the specific application where it is deployed. More precisely, first a backbone network (e.g., AlexNet, VGG16, ResNet) is chosen. Trained on large-scale generic object recognition datasets, such backbone architectures have already learned the invariant features applicable to all object recognition tasks and applications. Then, the baseline network is trained on a fine-grained dataset to tune the parameters according to the desired application.

The highly investigated object recognition applications, the corresponding datasets, and competitions are discussed in the following.

4.1. Autonomous Driving

Perception is one of the main components of Autonomous Vehicle (AV) navigation [238]. Success in path planning, control, localization and mapping of the vehicle depends on a robust visual understanding of the surrounding environment. In the past 20 years, there have been vast research efforts on AV-related perception problems such as object recognition (e.g., traffic lights, pedestrians, and other vehicles), object tracking, visual odometry/SLAM, and optical flow. Since the early research developments on autonomous vehicles, there has been a focus on annotated datasets for relevant object recognition tasks. Such datasets are collected either using drones (bird’s eye view images), traffic cameras, or by sampling frames from hours of videos recorded by onboard cameras of one or a fleet of vehicles. The range of tasks covered in AV perception datasets is truly extensive. In accordance with the scope of the paper, we focus on the object recognition statistics of datasets.

The majority of AV perception datasets are taken in street-view, being similar to the viewpoint of perception sensors mounted on AVs. While some of the datasets only offer annotated RGB images (e.g., CamVid [28], Cityscapes [55], Mapillary Vistas [177], D^2 -City [31], BDD100k [265]), a number of AV street-view datasets are multimodal, e.g., KITTI dataset [85], Apolloscape dataset [243], Waymo [221], nuScenes [31], Oxford-1000km [163], Lyft V5 [120]. These datasets are collected using a research vehicle equipped with a variety of sensors such as LiDAR, RADAR, RGB cameras, inertial measurement units, and driven in various locations, weather conditions, and time of the day. While incorporating valuable layered perception data, multimodal datasets are expensive to collect and bring about other difficulties such as correct integration, synchronization and calibration of sensors.

The main annotation choice for street-view datasets is 2D and 3D bounding boxes [31, 86, 188, 37] as it is more convenient for object tracking purposes [62]. Multimodal datasets usually offer a temporal correspondence between the annotated objects in different perception modules. Several datasets are annotated with rich pixel-wise segmentation masks instead of bounding boxes. CamVid [27] is the first dataset offering semantic segmentation. Apolloscape [243] provides semantic segmentation masks for 140k camera images captured in diverse traffic conditions. The CityScapes dataset [55] is the first dataset with significantly large numbers of pixel-wise labelled frames. It has been collected through 50 different cities depicting varied driving conditions. The Mapillary Vistas dataset [177] surpasses the amount and diversity of labelled data

compared to Cityscapes [88]. The virtual Semantic KITTI dataset [14] provides sequential images with depth information and pixel-wise labelled dense point cloud video frames. BDD-100k dataset [265] is focused on frame annotation and pixel-wise labelled dense point cloud video frames. Diversification of driving scenarios is another important factor with regards to street-view AV datasets. A number of datasets are collected with an emphasis on diverse weather and illumination conditions, e.g., D2 city [31], Mapillary Vistas [177], and BDD-100k dataset [265]. There are also datasets that are collected by ego-vehicles driven in various cities or even countries. The details of the most popular street-view datasets are provided in Table 8.

Dataset	Year	Location	Annotated frames	Number of Classes	Object Annotations	Highlights
KITTI [85]	2012	Karlsruhe, Germany	15k	8	200k 3D BB	Pioneer benchmark dataset for 3D object detection; multimodal
Cityscapes [55]	2016	50 cities in EU	25k	27	65k SM	Rich annotations provided; scene variability and complexity; depth information provided
BDD 100k [265]	2017	NY, SF	100k	40 Objects, 8 Lanes	1.8M BB	Diversified in location and weather conditions; instance segmentation masks provided for 10k images of the dataset
KAIST [50]	2018	Seoul	8.9k	3	308k BB	All-day time conditions (e.g., sunrise, morning, noon); multimodal
ApolloScape [243]	2018	4x China	144k	25 Objects 28 Lanes	70k 3D BB	Contains lane markings based on the lane colors and styles; instance level annotations are available
A*3D [188]	2019	Singapore	39k	7	230k 3D BB	Focused on pedestrian detection; high driving speed and low annotation speed
Argoverse [37]	2019	Miami, Pittsburgh	22k	15	993k 3D BB	Focused on 3D object tracking and motion forecasting; annotated HD semantic maps included
Automotive RADAR [170]	2019	Germany	500	7	3000 3D BB	RADAR data and object detection based on RADAR data
H3D [186]	2019	San Francisco	27k	8	1.1M 3D BB	Full-surround 3D multi object detection and tracking
nuScenes [31]	2019	Boston	40k	23	1.4M 3D BB	First dataset providing 3D dataset with attribute annotations; first to provide RADAR data; rich multimodal information
Waymo [221]	2019	3x USA	200k	4	9.9M BB, 12 3D BB	Diverse weather and geographic conditions in the US; multimodal
Mapillary Vistas [177]	2017	Global	25k	152	8M SM	Scene-parsing with instance-level object segmentation with a diverse geographic, weather, season and daytime extent
Lyft L5 [120]	2019	Paolo Alto	46k	9	1.3M 3D BB	Multimodal captured by a fleet of vehicles; an annotated LiDAR semantic map is provided
D ² -City [39]	2019	China	700k	12	50k BB	Sampled from dashcam video sequences; Bounding box annotations

Table 8: Popular street-view autonomous driving datasets. BB = Bounding Boxes, SM = Segmentation Masks.

A number of datasets cover specific recognition problems in the AV domain e.g., pedestrian detection, car detection and traffic sign detection. Pedestrian detection has been heavily investigated in autonomous driving and other applications such as video surveillance. Such datasets could be divided into two categories. ‘Person’ datasets contain people instances in all kinds of poses while ‘Pedestrian’ datasets only consider upright people (walking and standing) typically viewed from more restricted viewpoints but often containing motion information and detailed labelling [80]. The famous pedestrian detection datasets are described in Table 9.

Dataset	Year	Number of Cities	Number of Images	Number of Pedestrians	Highlights
CityPersons [270]	2017	27 cities in EU	5000	35016	Built on top of the Cityscapes dataset
INRIA [58]	2005	-	614	902	Occlusion labels included
Caltech [64]	2009	1	250,000	2300	Temporal correspondence and occlusion labels included; sampled from 10 hours of video
MIT [184]	2000	-	1800	1800	Labelled using the LabelMe annotation tool
EuroCity [26]	2018	31 cities in EU	47,000	238,000	Largest pedestrian detection dataset to date
NightOwls [178]	2018	7	32	55,000	Pedestrian detection at night time; attribute annotations include pose, occlusion, and height
Daimler [66]	2009	1	21,790	56,492	Occlusion attributes provided; monocular images

Table 9: Pedestrian detection datasets. Number of images does not include unannotated images. Unique pedestrians are considered for the number of pedestrians.

Bird’s eye view datasets are captured using cameras mounted on drones, or from traffic signal lights or buildings in suitable vantagepoints over vehicles and pedestrians. Examples of such datasets are INTERACTION [267], NIGSM [54], HighD [125] and KITTI bird’s eye view. Bird’s eye datasets are usually collected with the purpose of vehicle and pedestrian behavior prediction, however since many of them are annotated, they could also be studied under the object recognition dataset category. A list of the popular bird’s eye view datasets with object annotations can be found in Table 10.

A number of challenges have also been held to boost research in recognition tasks. Table 13 provides the details of the the predominant AV-related object recognition datasets.

4.2. Medical Imaging

The promotion of computer vision applications in the medical field has made a significant improvement to healthcare industry all over the world. Artificial intelligence (AI) in medicine can be generally classified into two categories, virtual and physical [6]. A virtual dataset is made up of images taken by medical equipment such as X-ray, and computed tomography scanner. These datasets aim to increase the accuracy and

Dataset	Year	Location	Road span/Area	Size of data	Highlights
NGSIM [54]	2005	USA	500-640m Span of road	90 min	Video cameras attached to the adjacent buildings; speed levels more than 75km/h are not included in the dataset
High D [125]	2017	Germany	420m Span of road	16.5 hours	Drone based dataset with five scenario description layers, the first 3 layers include static scenario description, 4th layer includes dynamic description, 5th layer includes environment conditions
The inD [22]	2017	Germany	Altitude 100m 80x40 meters to 140x70 meters	10 hours of video recording	The dataset includes more than 11,500 road users including pedestrians, bicyclists and vehicles at intersections
INTERACTION [267]	2019	USA, China, Bulgaria, Germany	n/a	365min+433min+133min+60min	Data collected from drones and traffic cameras; Multimodal
AU-AIR [25]	2019	Aarhus, Denmark	Flight altitude (5m to 30m) and camera angle 45 to 90 degree	2 hours	Multimodal; compare and contrast of natural and aerial images in the context of object detection

Table 10: Bird’s eye view datasets

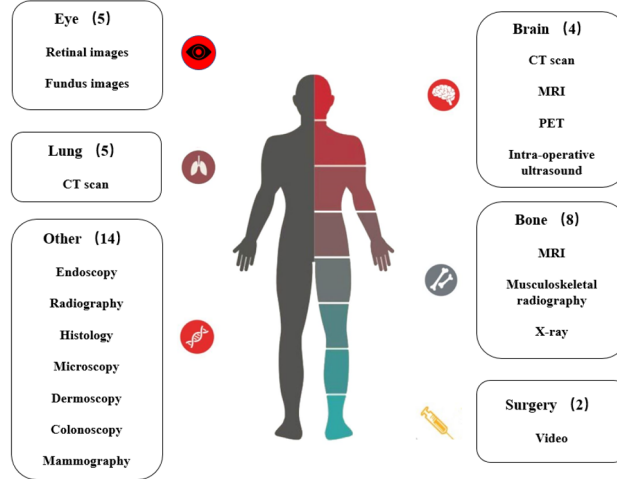


Figure 5: Different types of medical images included in datasets. The number of available datasets related to each organ is shown in parentheses.

speed of certain diseases diagnosis. Thus, the doctors can plan treatment in advance and treat the patients effectively. A physical dataset consists of videos filmed during surgery procedures. The goal of a physical dataset is to promote the implementation of computer and robots to surgery.

Numerous datasets and benchmarks have been produced using various medical imaging equipment for different organs. Figure 4 shows different types of medical images included in mentioned datasets, and a summary of major datasets and challenges could be found in Tables 11 and 12 respectively.

4.3. Face Recognition

Face Recognition (FR) is one of the highly investigated object recognition tasks with datasets and algorithms dating back to the early 1990’s. The main two face recognition tasks are face verification and face identification. The former determines if a pair of images correspond to the same human subject (e.g., used in access control systems). This task has been widely studied with algorithms outperforming human recognition rates [119] on the benchmark datasets such as LFW [108]. For face identification, given two gallery and probe face image sets, the task is to rank the images in the gallery set based on their similarity to each image in the probe set. Forensic identification, watchlist identification, and de-duplication (e.g., mug shot repositories, ID card databases) are among the applications of face identification paradigm [124]. In the past three decades, there has been a steady expansion of the FR datasets as evolving algorithms reached near-perfect accuracy levels on previous datasets. The early face recognition datasets offered laboratory-controlled samples with limited scale, lighting condition, and background clutter. FERET [189], CMU-PIE [214], and the AR Face database [166] are among the prominent datasets of this era. A complete list of early milestone datasets is provided in [2]. After the introduction of shallow learning methods, which used hand-engineered descriptors such as SIFT and HOG, these datasets became saturated, and a new era of datasets containing unconstrained images emerged to challenge the state-of-the-art. LFW was among the

Challenge/Benchmark	Year	Task	Dataset	Metric	Task(s)
CVPR 2018 - Video Segmentation Challenge [115]	2018	Video Segmentation	-	mAP & IoU	Segmentation of movable object from video frames
CVPR 2018 - Berkeley DeepDrive challenges [18]	2018	Road Object Detection & Drivable Area Segmentation & Domain adaptation	BDD 100K dataset	AP & IoU	Object detection, segmentation and tracking
nuScenes 3D detection challenge [31]	2019	3D model generation	nuScenes dataset	mAP & TP	Object detection with attributes using 3D bounding boxes
Lyft 3D Detection for Autonomous Vehicle [162]	2019	Object detection	Lyft Level 5 dataset	IoU	Object detection with 3D bounding boxes over semantic maps.
NightOwls Pedestrian Detection Challenge [178]	2019	Pedestrian detection	NightOwls dataset	Standard average missing rate	Pedestrian detection at night time
D ² -City Detection Domain Adaptation Challenge [62]	2019	Object detection & Domain adaptation	Image-Net & BDD 100K datasets	AP & IoU	Object detection in diverse conditions with a focus on transfer learning to new unseen environments
WIDER Face & Person Challenge	2019	Pedestrian detection	WIDER dataset	mAP & IoU	Detection of pedestrians and cyclist in unconstrained environment.
CVPR 2019 - Beyond Single-frame Perception [7]	2019	3D object detection	-	mAP & IoU	Object detection on Lidar data
The KITTI 3D Object Evaluation Benchmark [84]	2017	Object detection	KITTI dataset	precision-recall curve & AP	Object detection using 3D bounding boxes
Caltech Pedestrian Detection Benchmark [64]	2012	Pedestrian detection	Caltech Pedestrian dataset	IoU	Pedestrian detection
The KITTI 2D Object Evaluation Benchmark [83]	2012	Object detection & Object Orientation	KITTI dataset	precision-recall curve & AP & average orientation similarity	Object detection and orientation estimation

Table 11: AV-related object recognition and scene understanding challenges.

first datasets to offer unrestricted images in the wild, varying in illumination levels, subject pose, and facial expression. Due to its increased difficulty compared to the laboratory-controlled datasets, LFW remained the main dataset for benchmarking FR algorithms for several years. [249, 129, 17, 69] are among other datasets in this era. The third generation of FR datasets were increased in scale to fuel the data-hungry deep FR algorithms. Similar to other object recognition paradigms, it was shown that the accuracy of deep FR networks benefits from the availability of large-scale training data [275]. The early large-scale datasets were private and hence could neither be used as fair performance benchmarks nor as training sets for deep models developed in the academia. Facebook’s DeepFace [228], Google’s FaceNet [208], and the DeepID series [224, 223, 222], are the major models trained on private datasets of 4M, 500M, and 0.2M images respectively. Such models reached near perfect accuracies on benchmarks such as LFW and YTF, hence calling for more challenging benchmarks. To this end, public datasets were gradually introduced to both enable large-scale training of academic deep models (e.g., CASIA-WebFace [264], MS-Celeb-1M [93], VGGFace2 [36]) and also to create a fair performance evaluation medium (e.g. [124, 247, 168] and MegaFace [119, 174] benchmarks).

FR datasets expand either in breadth or depth. Broad datasets are focused on increasing the number of subjects and are suitable for low-shot learning purposes [174, 204, 93], while deep datasets increase their size by containing more images per subject [36, 41, 185], therefore allowing for intra-class variations resulted from age, pose, and expression variations. Some datasets also offer additional annotations such as facial keypoints, age, pose, craniofacial features, and binary attributes. The prominent datasets from the second and third eras are listed in Table 14.

Most FR datasets are collected using internet sources and consist of celebrity faces in formal occasions. This limits the studied demographic. Data bias exists with regards to gender, race and age. It has been shown that commercial and state-of-the-art algorithms perform best on Caucasians while having the most error rates on African faces for the verification task [244]. To this end, there have been recent efforts to correct the recognition bias by decreasing class imbalance [244, 117] and including diversity-based features [169].

Since face verification and face identification are two different matching problems (one-to-one versus one-

Dataset	Size	Year	Targeted disease/organ	Content	Challenge	Description
NLM's MedPix Dataset [173]	59000 images	-	-	Integrated images	no	A free online dataset contains more than 12000 patient cases
STARE Database [90]	400 cases	-	Eye	retinal images	no	Blood vessel segmentation images
SMIR [164]	350425 images	-	-	CT scans	yes	51 subjects of whole-body postmortem CT scans
EchoNet-Dynamic [59]	10030 images	2020	Heart	Echocardiographic video frames	yes	An expert labeled dataset for the study of cardiac motion and chamber size.
Atlas of Digital Pathology [107]	17668 images	2020	Radiological diagnosis	Histological patch images	yes	Images of different organs with 57 types of hierarchical tissue annotated
COVID-CT Dataset [267]	349 images	2020	COVID19	CT scans	no	-
SARAS-ESAD Dataset [57]	22601 frames	2020	Prostatectomy procedure	Video frames	yes	A dataset of videos showing the full prostatectomy procedure by surgeons
The StructSeg 2019 Dataset[144]	120 cases	2019	Radiotherapy planning	CT scans	yes	A dataset for the treatment of cancers
ODIR-5K [181]	5000 images	2019	Eye	fundus photographs	yes	Fundus images taken by various cameras with different resolutions
DRIVE [219]	400 cases	2019	Eye	Retinal images	yes	Images of 400 different patients between 25-90 years of age.
The RSNA Brain Hemorrhage CT Dataset [79]	874035 images	2019	Brain Hemorrhage	CT scans	yes	Images gathered from 2 medical societies and 60 neuroradiologists
The KiTs19 Challenge Dataset [101]	300 cases	2019	Kidney tumor	CT scans	yes	A dataset of multi-phase CT imaging with segmentation masks
SegTHOR [132]	60 scans	2019	Lung	CT scans	No	A dataset focused on the segmentation of organs at risk in the thorax
The EAD Challenge Dataset [4]	2700 images	2019	Hollow organs	Endoscopic video frames	yes	Images collected from 6 different data centers
Oasis Brains Dataset [133]	~1000 cases	2019	Brain	MRI & PET images	no	A dataset collected over 30 years
CheXpert [110]	224316 images	2019	Chest	Chest radiographs	yes	A dataset labeled by an automatic labeler
LERa [141]	182 patients	2019	Musculoskeletal disorder	Radiographs	yes	Images of hip, foot, ankle and knee of patients for the study of musculoskeletal disorders
CAMEL colorectal adenoma Dataset [259]	177 cases	2019	Cancer	Histology images	no	A dataset for segmentation of cancerous parts in organ
BACH Dataset [9]	430 images	2019	Breast cancer	Microscopy & whole-slide images	yes	Microscopy images labelled by 2 experts
MRNet [21]	1370 patients	2018	Knee	MRI	yes	A dataset for autonomous MRI diagnosis
The REFUGE Challenge Dataset [182]	1200 images	2018	Glaucoma	Fundus photographs	yes	The dataset was collected using two types of devices
MURA [193]	40561 images	2018	Bone	musculoskeletal radiographs	yes	A manually labeled dataset by board-certificated Stanford radiologists, containing 7 body types: finger, hand, elbow, forearm, humerus, wrist and shoulder
Calgary-Campinas Public Brain MR Dataset [218]	167 scans	2018	Brain	MRI	no	A dataset for analysis of brain MRI
HAM 10000 Dataset [235, 53]	10015 images	2018	Skin lesions	Dermatoscopic images	yes	A multi-modal and multi-population dataset
NIH Chest X-ray Dataset [246]	100000 images	2017	Chest	X-ray images	no	A dataset of x-ray images
609 Spinal Anterior-posterior X-ray Dataset[251]	609 images	2017	Spine	X-ray images	No	Each vertebra was located by a landmark and the landmark is used to calculate Cobb angles.
RESECT [257]	23 patients	2017	Cerebral Tumor	MRI & intra-operative ultrasound	yes	A dataset of homologous landmarks
Cancer Digital Slide Archive [248]	-	2017	Cancers	Glass slides of histologic images	no	High resolution detailed images of tissue microenvironments and cytologic details
Cholec80 [236]	80 videos	2016	Surgery	Video frames	no	A dataset containing 80 videos of surgeries performed by 13 different surgeons
CSI 2014 Vertebra Segmentation Challenge Dataset [263]	10 scans	2016	Spine	CT scan	yes	Entire thoracic and lumbar spine were covered by the images. The in-plane resolution is from 0.31 to 0.45mm. The slice-thickness is 1mm or 2mm.
CRCHistoPhenotypes - Labeled Cell Nuclei Data[215]	100 images	2016	Cell	Histology images	no	100 H&E stained histology images of colorectal adenocarcinomas
Multi-Modality Vertebra Dataset [33]	20 cases	2015	Vertebra	MRI & CT scan	no	The 3D vertebra centre location and orientation are annotated.
CVC colon DB [19]	1200 frames	2012	colon & rectum	Colonoscopy video frames	no	The dataset's region of interest has been annotated. The video frames were specifically chosen for maximum visual distinction among them.
LIDC-IDRI Database [10]	1018 cases	2011	Lung nodule	CT scans	yes	A database created by 7 academic centers and 8 medical imaging companies
Computed Tomography Emphysema Dataset [217]	115 slices	2010	COPD	CT scans	no	High-resolution CT scans
DIARETDB1 [118]	89 images	2007	Diabetic retinopathy	fundus photographs	no	A database for benchmarking the detection of diabetic retinopathy
ELCAP Public Lung Image Database [65]	50 sets	2003	Lung	CT scans	no	50 low-dose documented CT scans for lungs containing nodules
The Digital Database for Screening Mammography [98, 99]	2620 cases	1998	Breast	Mammography images	no	The database has the function for user to search classes among normal, benign and cancer.

Table 12: Medical imaging datasets.

to-many), different evaluation metrics are used to assess the algorithms in these paradigms. Face verification algorithms are usually evaluated using the ROC curve. For a given test set of image pairs, a ROC curve can be drawn by changing the threshold above which the output confidence score is considered a matched pair. Each point on the ROC curve corresponds to a True Positive Rate (TPR) on the vertical and a False Positive Rate (FPR) on the horizontal axis for a threshold t (the independent variable). The TPR is the fraction of detected face matches that correctly exceed t , and FPR is the fraction of detected face matches that falsely exceed the threshold. The area under the ROC curve (AUC) or the TPR value at a fixed FPR can then be calculated as a measure of the verification accuracy. Another face verification metric is LFW's *estimated mean accuracy* [108], which is the percentage of correct binary classifications. The Cumulative Match Characteristic (CMC) curve is the most common evaluation metric used for face identification. Given a set of probe images and a gallery, the CMC curve can be drawn by ranking the images in the gallery according to predicted similarity scores for each of the probes. For each point on the CMC curve, the vertical axis value corresponds to the percentage of the probe images for which the algorithm has found the corresponding gallery mate within the top- k ranked predictions, and the horizontal axis value

Challenge	Year	Task	Dataset	Metric	Targeted area
SARAS [56]	2020	Detection	SARAS-ESAD Dataset	mAP	AI integrated minimally invasive surgery; Detection of surgeon's actions
REFUGE [182]	2020	Detection Segmentation	REFUGE Challenge Dataset	-	Automated segmentation and detection of glaucoma
StructSeg [145]	2019	Segmentation	The StructSeg 2019 Dataset	DSC & 95% Hausdorff Distance	Lung cancer and nasopharynx cancer; Gross target volume and organs at risk
DRIVE [219]	2019	Segmentation	DRIVE	Overall prediction accuracy and s	Screening programs for diabetic retinopathy; Diagnosis of hypertension and computer-assisted laser surgery
ODIR [181]	2019	Classification	ODIR-5K	Precision, accuracy and dice similarity	AI in retinal image analysis
RSNA Intracranial Hemorrhage Detection [192]	2019	Detection	The RSNA Brain Hemorrhage CT Dataset	Weighted multi-label log loss	Detection of acute intracranial hemorrhage and respective subtypes
KITS [237]	2019	Segmentation	The KiTS19 Challenge Dataset	PROC	Kidney tumor semantic segmentation
EAD [5]	2019	Classification Detection Segmentation	The EAD Challenge Dataset	average Dice coefficient	Diagnosis and treatment for diseases in hollow organs
AASCE [147]	2019	Regression	AASCE Challenge Dataset	Symmetric mean absolute percentage error	Automated spinal curvature estimation and correction of error from x-ray images
CurIOUS [201]	2019	Registration	RESECT	Threshold Jaccard Index and normalized multi-class accuracy	Employment of AI in surgery
ISIC [35]	2018	Classification	The HAM 10000	mAP, IoU, Dice coefficient, Jaccard Index, F2 score and deviation score.	Automated diagnosis of melanoma
Data Science Bowl [34]	2018	Classification	-	IoU	Detection of nucleus
ICIAR [8]	2018	Classification Segmentation	BACH	Mean target registration errors	Early diagnosis of breast cancer
LUNA [111]	2016	Classification Detection	LIDC-IDRI	Kappa score, F1 score and AUC	Automatic detection of lung nodule algorithms
SARAS [56]	2015	Segmentation	-	Average Dice coefficient	Detection of dental caries

Table 13: Medical imaging challenges and benchmarks.

Dataset	Year	Number of Subjects	Number of Images	Additional Information	Highlights
VGGFace [185]	2015	2,622	2.6 M	A	Large-scale celebrity recognition with high intra-class variations
VGGFace2 [36]	2018	9,131	3.31M	A, P	Diversified pose, age, and ethnicity of celebrity faces
LFW [108]	2007	5,749	13,233	-	The first unconstrained FR dataset
MegaFace [174]	2016	672,052	4.7 M	-	Raised difficulty by including 1 M distractors; non-celebrity subjects
YTF [249]	2011	1,595	3,425 V	-	Designed for face verification in videos; same format as LFW
CASIA-WebFace [264]	2014	10,577	494,414	-	First publicly available large-scale FR dataset
IJB-A [124]	2015	500	5,712	BB, KP	Manually verified bounding boxes for face detection, nose and eye keypoints included
MS-Celeb-1M [93]	2016	100,000	10 M	-	Celebrity identification dataset and benchmark with a linked celebrity knowledge base
Pubfig [129]	2009	200	60,000	A, E, G, P, BB	73 automatically generated attributes provided; same format as LFW
CelebA [160]	2015	10,177	202,599	KP	Designed for face attribute prediction in the wild; 40 binary attributes included
DiF [169]	2019	-	0.97 M	A, P, S, BB, KP	Quantitative facial features included to reduce recognition bias across different demographics
IMDB-Face [205]	2015	100,000	460,723	A, G	Age and gender prediction on a set of celebrities collected from IMDB
UMDFaces [11]	2016	8,501	367,920	A, P, G, BB, KP	Detailed human-verified attributes and annotations
IJB-B [247]	2015	1,845	21,798	A, G, P, S	A superset of IJB-A with additional occlusion and illumination annotations
IJB-C [168]	2018	3,531	31,334	A, G, P, S	An improvement upon IJB-B with a focus on diversifying the geographic coverage of subjects
FaceScrub [104]	2014	695	141,130	G	A broad dataset of movie celebrities gathered from IMDB
CACD[41]	2014	2,000	163,446	A	Images include age variations for each subject for cross-age face recognition and retrieval; only 200 subjects are manually annotated

Table 14: Well-known face recognition datasets. Abbreviations in the table: Occlusion (O), Pose (P), Age (A), Expression (E), Skin color (S), Gender (G), Bounding Boxes (BB), Keypoints (KP), V (video).

represents the rank k , varied between one and the total number of images in the gallery. The MS-Celeb-1M is the only major face identification benchmark that uses a precision-coverage curve instead of the CMC curve [93]. A list of metrics used in common FR benchmarks is provided in [167].

4.4. Remote Sensing

The increased access to aerial images captured by satellites or UAVs has paved the way for a wide series of applications such as traffic flow management, precision agriculture, natural disaster response, and map extraction. In order to achieve a scaled utility, such applications call for automated detection of map features such as urban infrastructure, vegetation, and other points of interest.

Remote sensing datasets usually are limited in size and the number of covered classes compared to object recognition datasets in ground-level view. Vehicles, roads, and buildings are among the highly investigated object classes due to their significance in a wide range of applications such as traffic management, surveillance, urban planning and modeling population dynamics [60, 172]. In recent years, a number of general-purpose object recognition datasets have also emerged, covering more classes [48, 279, 277] with a

few exceptions such as xView [131] and fMoW [52], these datasets still lack the scale required for training deep models from scratch. An overview of some of the popular remote sensing datasets is provided in Table 15.

While many of the datasets are scraped from Google Earth (e.g., NWPU [48], DOTA [254], TAS [100]) and thus lack multispectral data and temporal views, some datasets utilize 4-band or 8-band images to include a wider range of electromagnetic spectrum, especially the near infrared channel [196].

The dominant choice of annotation for remote sensing datasets is bounding boxes. Since objects appear in different orientations when looked from above, remote sensing images contain high levels of intra-class aspect ratio variations. Thus, many of the remote sensing object detection datasets offer rotated bounding box annotations to alleviate this problem [254, 196]. Several datasets have also been proposed for semantic segmentation purposes, e.g., the TorontoCity [245], ISPRS 2D semantic labeling [209], DeepGlobe+Vivid [60] datasets. A number of competitions have also been held to promote research in aerial object recognition. A list of important remote sensing challenges is provided in Table 16.

Dataset	Year	Annotation	Size	Spatial Resolution (cm per pixel)	Description
SpaceNet C.1&C.2	2019	685,000 buildings	5,555 km^2	30-50	Building footprints annotated using polygons; five cities
SpaceNet C.3	2019	8,676 km road	5,555 km^2	30-50	Road centerlines labeled based on the OpenStreetMap scheme
COWC [172]	2016	32,716 vehicles	-	15	Car detection dataset gathered from six cities in North America and Europe; cars annotated with points on centroids
xView [131]	2018	1M objects	1,400 km^2	30	Large-scale object overhead object detection dataset with bounding box annotations
FMoW [52]	2017	132,700 objects	1M	-	Temporal image sequences from over 200 countries with the purpose visual reasoning about location, time, and sun angles; bounding box annotations
NWPU-RESISC45 [48]	2017	31,500 scenes	31,500	20-3000	Aerial scene classification dataset with variations in spatial resolution, illumination, object pose, occlusion
TorontoCity [245]	2016	400,000 buildings	712 km^2	10	RGB and LiDAR Aerial imagery of the greater Toronto area augmented with and street-view stereo and LiDAR
DOTA [254]	2018	188,282 objects	2,806	-	Rotated bounding box annotations verified by expert annotators; 15 common classes
TAS [100]	2008	1,319 vehicles	30	-	An early annotated remote sensing dataset from collected from google earth; bounding box annotations
DLR3K [154]	2013	3,472 vehicles	20	13	Rotated bounding boxes with additional orientation annotations
NWPU VHR-10 [47]	2016	3,775 objects	715	50-200	generic object detection dataset with 10 classes; bounding box annotations
LEVIR [279]	2018	11,000 objects	22,000	20-100	Bounding box annotations provided for airplanes, ships, and oilpots
VEDAI [196]	2016	3,600 vehicles	1,210	12.5	Small vehicle detection consisting of 9 vehicle classes; rotated bounding box annotations
UCAS-AOD [277]	2015	6,000 objects	910	-	Rotated bounding box annotations; vehicle and airplane detection; scraped from Google Earth
AID[255]	2016	10,000 scenes	10,000	50-800	Aerial scene classification with 30 classes

Table 15: Remote sensing object detection datasets. Dataset size is the number of images unless states otherwise.

Challenge	Year	Dataset size	Number of Classes	Evaluation Metric	Task
SpaceNet C.1&C.2[67]	2019	5,555 km^2	2	F_1 score	Building footprint detection in 5 cities
SpaceNet C.3[67]	2019	5,555 km^2	1	Average Path Length Similarity	Road network extraction
FMoW [52]	2017	1 M	63	F_1 score	Object classification
DSTL [116]	2017	57	10	IoU	Semantic segmentation
NWPU- RESISC45 [48]	2017	31,500	45	Accuracy	Scene classification
DIUx xView [131]	2018	1,400 km^2	60	IoU	Object detection
DeepGlobe [60]	2018	10,000	7 *	IoU, F_1 score	Building segmentation, road extraction; land cover classification

Table 16: Remote sensing challenges. * the number of classes for the land cover classification task.

4.5. Species Recognition

Numerous fine-grained datasets have been created for species recognition tasks. Although there are many images of various pet animal categories available online, datasets including other categories e.g., wild animals are very sparse. Therefore, many efforts have been made to collect images of diverse species in

their natural habitats [105, 226, 89]. Another challenge with regards to fine-grained species recognition is the required expertise knowledge for fine-grained classification annotations, which makes crowdsourcing less practical. However, crowdsourcing workers have also been employed to help with bounding box or keypoint annotations [121, 68]. The popular species recognition datasets are described in Table 17. Species recognition competitions are also held for some of these datasets. iNaturalist is an annual Kaggle-hosted competition offering a classification task of a subset of the main dataset and is evaluated using a combination of ILSVRC’s top-k classification evaluation criteria. LifeCLEF is another annual competition with a plant identification task that is based on the PLANTCLEF dataset and is evaluated by the top-1 accuracy criterion.

Dataset	Number of Images	Number of Classes	Number of Annotations	Year	Challenge	Description
Flower 102 [179]	8,189	103	8,189 SM	2008	No	Flower recognition dataset of 103 flower categories common in the United Kingdom
Caltech-Birds 2011[241]	11,788	200	11,788 BB	2011	No	15 part locations and 28 attributes for each bird
Stanford Dogs [122]	22,000	120	22,000 BB	2011	No	Single-object per image dataset for dog breed recognition
F4K [23]	27,370	23	27,370 CL	2012	No	Fish recognition dataset annotated by following marine biologists
Snapshot Serengeti [226]	1.2 m	61	406,433 CL, 150,000 BB	2014	No	Wild animal classification dataset gathered using 225 camera-traps in Serengeti national park in Africa
NABirds [106]	48,562	555	48,562 BB	2015	No	Expert-curated dataset of North American birds; 11 bird parts annotated in every image
PlantCLEF [89]	434,251	10,000	10,000 CL	2015	Yes	Plant classification dataset gathered in the Amazon rainforest
iNat [105]	675,175	5,089	561,767 BB	2017	Yes	Manually collected dataset of 13 super-class and 5k sub-class species; organized in a hierarchical taxonomy; highly imbalanced
Dogs-in-the-Wild [220]	300,000	362	300,000 CL	2018	No	A large dataset for dog breed classification in natural environments
AnimalWeb [121]	21,900	334	198k KP	2019	No	Hierarchically categorized dataset for animal face recognition with 9 keypoint annotations per face
IP102[252]	75,000	102	75,000 CL, 19,000 BB	2019	No	Hierarchically categorized dataset for insect pest recognition

Table 17: Species recognition datasets

4.6. Clothing Detection

Fashion is another milieu that has attracted a lot of attention in the computer vision community. There are two vision tasks that are most popular: 1) retrieval of clothing images similar to an item in a given image, in the same manner as the recommendation systems used in cloths shopping websites, and 2) parsing individual fashion items where the goal is to identify the exact model of a clothing item in an image taken by the customer and retrieve the same product or similar products from the shopping floor [272]. To this end, the created datasets usually offer multiple clothing attributes for each item in the form of binary tags (e.g., has v-shaped collar, is red), and some of them also offer clothing pairs, one from an item in the shopping floor and the other taken by a customer in an unconstrained condition to facilitate the retrieval procedure. Although mere recognition is usually not the main purpose of clothing datasets, some of the clothing datasets offer fine-grained detection/segmentation annotations, which can be used for accurate recognition purposes. Clothing recognition remains a challenging task mainly due to the non-rigid structure of cloths that cause unpredictable deformations in various body poses. Other challenges include occlusions, sensitivity to lighting conditions, and inconsistencies between commercial and customer images [82]. The prominent clothing datasets with annotated instances are discussed in Table 18. The “annotated clothing instances number”

Dataset	Number of Images	Classes	Number of Annotated Clothing instances	Annotation Type	Year	Challenge/Benchmark	Number of Attributes
DARN[109]	182,000	20	182,000	BB	2015	No	9
Street2Shop[123]	404,000	11	20,357	BB	2015	No	-
DeepFashion[159]	800,000	50	180,000	KP	2016	Yes	5
ModaNet[272]	55,000	13	240,000	BB, SM	2018	No	-
FashionAI[278]	324,000	41	324,000	KP	2018	No	68
Deepfashion2 [82]	491,000	13	801,000	BB, SM, KP	2019	Yes	4

Table 18: Clothing detection datasets

column in the table shows the number of cloths having annotations that could be used for detection purposes.

As mentioned before, some datasets have unannotated images from shopping retailers that can only be used for image retrieval and not for detection. Deepfashion 2 is the richest annotated clothing dataset, which also comes with an evaluation benchmark. The benchmark tasks are cloth detection, landmark estimation, and segmentation, which are evaluated based on analogous tasks in MS COCO challenges. The cloth detection and segmentations tasks are evaluated using AP for bounding boxes and segmentation masks respectively. The landmark estimation task, which is the localization of clothing keypoints, is evaluated by AP for the annotated keypoints (same as MS COCO keypoint detection).

5. Conclusion and Future Work

In this paper, a detailed analysis on the milestone datasets in generic object recognition was provided. This review also focused on six popular applications namely autonomous driving, medical imaging, face recognition, remote sensing, species detection, and clothing detection. With a focus on the most recent datasets generated in the deep learning era, details such as dataset size, annotation type, and individual descriptions are mentioned, and the challenges faced by the community in data collection, annotation, and algorithm design are also discussed. Our concentration on the study of the publicly available datasets and cataloguing the researchers' efforts involved in generation and optimizing them were in view of the fact that the deep learning community generally believes that the most critical upcoming challenge in way of improving the utility of deep learning and other data-driven methods involves introducing and augmenting comprehensive and flawless training datasets.

In addition to the datasets, the major generic and application-specific object recognition competitions were investigated, with information on the adopted evaluation metrics, covered tasks, and dataset statistics. A comprehensive overview of the evaluation criteria frequently used in the competitions and benchmarks is also provided to more clarify the evaluation processes used for assessing the performance of different tasks. This will provide the curious reader with clues about the reasons for success of the approaches taken and an insight to where to begin in their own projects.

In conclusion, this review is an effort to help researchers shorten the process of finding the appropriate training and testing mediums for their desired applications and to shed light upon promising data collection directions as the state-of-the-art models inevitably saturate the existing datasets.

Acknowledgements

We would like to acknowledge the financial support received by Aria Salari for this survey from Mathematics of Information Technology and Complex Systems (MITACS) under the Accelerate Program Internship no. IT16412 and Vancouver Computer Vision.

References

- [1] , . IEEE Xplore Full-Text PDF:. URL: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp={\&}arnumber=8241865>.
- [2] Abate, A.F., Nappi, M., Riccio, D., Sabatino, G., 2007. 2D and 3D face recognition: A survey. *Pattern Recognition Letters* 28, 1885–1906. doi:10.1016/j.patrec.2006.12.018.
- [3] Achantay, R., Hemamiz, S., Estraday, F., Süssstrunk, S., 2009. Frequency-tuned salient region detection. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009 , 1597–1604doi:10.1109/CVPRW.2009.5206596.
- [4] Ali, S., Zhou, F., Daul, C., Braden, B., Bailey, A., Realdon, S., East, J., Wagnières, G., Loschenov, V., Grisan, E., Blondel, W., Rittscher, J., 2019a. Endoscopy artifact detection (EAD 2019) challenge dataset , 1–13doi:10.17632/C7FJBXCGJ9.1.
- [5] Ali, S., Zhou, F., Daul, C., Loschenov, M., 2019b. EAD 2019. URL: <https://ead2019.grand-challenge.org/>.
- [6] Amisha, Malik, P., Pathania, M., Rathaur, V.K., 2019. Overview of artificial intelligence in Medicine. *Journal of Family Medicine and Primary Care* 8, 2328–2331. doi:10.4103/jfmpc.jfmpc_440_19.
- [7] Apolloscape, 2019. CVPR 2019 WAD Beyond Single-frame Perception Challenge. URL: <http://wad.ai/2019/index.html>.
- [8] Araújo, T., Aresta, G., Eloy, C., António, P., Aguiar, P., 2018. ICIAR 2018. URL: <https://iciar2018-challenge.grand-challenge.org/>.

- [9] Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S.S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., Fernandez, G., Zeineh, J., Kohl, M., Walz, C., Ludwig, F., Braunewell, S., Baust, M., Vu, Q.D., To, M.N.N., Kim, E., Kwak, J.T., Galal, S., Sanchez-Freire, V., Brancati, N., Frucci, M., Riccio, D., Wang, Y., Sun, L., Ma, K., Fang, J., Kone, I., Boulmane, L., Campilho, A., Eloy, C., Polónia, A., Aguiar, P., 2019. BACH: Grand challenge on breast cancer histology images. *Medical Image Analysis* 56, 122–139. doi:10.1016/j.media.2019.05.010.
- [10] Armato, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., Van Beek, E.J., Yankelevitz, D., Biancardi, A.M., Bland, P.H., Brown, M.S., Engelmann, R.M., Laderach, G.E., Max, D., Pais, R.C., Qing, D.P., Roberts, R.Y., Smith, A.R., Starkey, A., Batra, P., Caligiuri, P., Farooqi, A., Gladish, G.W., Jude, C.M., Munden, R.F., Petkovska, I., Quint, L.E., Schwartz, L.H., Sundaram, B., Dodd, L.E., Fenimore, C., Gur, D., Petrick, N., Freymann, J., Kirby, J., Hughes, B., Vande Castele, A., Gupta, S., Sallam, M., Heath, M.D., Kuhn, M.H., Dharaia, E., Burns, R., Fryd, D.S., Salganicoff, M., Anand, V., Shreter, U., Vastagh, S., Croft, B.Y., Clarke, L.P., 2011. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics* 38, 915–931. doi:10.1118/1.3528204.
- [11] Bansal, A., Nanduri, A., Castillo, C., Ranjan, R., Chellappa, R., 2016. UMDFaces: An Annotated Face Dataset for Training Deep Networks. *IEEE International Joint Conference on Biometrics, IJCB 2017 2018-Janua*, 464–473. URL: <http://arxiv.org/abs/1611.01484>.
- [12] Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., Katz, B., 2019. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 1–11 URL: <https://objectnet.dev>.
- [13] Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: Speeded up robust features , 404–417.
- [14] Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J., 2019. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences URL: <http://arxiv.org/abs/1904.01416>.
- [15] Bell, S., Upchurch, P., Snavely, N., Bala, K., 2013. OPENSURFACES: A richly annotated catalog of surface appearance. *ACM Transactions on Graphics* 32. doi:10.1145/2461912.2462002.
- [16] Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., 2007. Greedy layer-wise training of deep networks, in: *Advances in neural information processing systems*, pp. 153–160.
- [17] Berg, T.L., Berg, A.C., Edwards, J., Maire, M., White, R., Teh, Y.W., Learned-Miller, E., Forsyth, D.A., 2004. Names and faces in the news. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2*. doi:10.1109/cvpr.2004.1315253.
- [18] Berkeley Deep Drive, 2018. CVPR 2018 - Berkeley DeepDrive challenges.
- [19] Bernal, J., Sánchez, J., Vilarino, F., 2012. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition* 45, 3166–3182. doi:<https://doi.org/10.1016/j.patcog.2012.03.002>.
- [20] Beumier, C., Achery, M., 2000. Automatic 3D face authentication. *Image and Vision Computing* 18, 315–321. doi:10.1016/S0262-8856(99)00052-9.
- [21] Bien, N., Rajpurkar, P., Ball, R.L., Irvin, J., Park, A., Jones, E., Bereket, M., Patel, B.N., Yeom, K.W., Shpanskaya, K., Halabi, S., Zucker, E., Fanton, G., Amanatullah, D.F., Beaulieu, C.F., Riley, G.M., Stewart, R.J., Blankenberg, F.G., Larson, D.B., Jones, R.H., Langlotz, C.P., Ng, A.Y., Lungren, M.P., 2018. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Medicine* 15, 1–19. doi:10.1371/journal.pmed.1002699.
- [22] Bock, J., Krajewski, R., Moers, T., Runde, S., Vater, L., Eckstein, L., 2019. The inD Dataset: A Drone Dataset of Naturalistic Road User Trajectories at German Intersections .
- [23] Boom, B.J., Huang, P.X., Beyan, C., Spampinato, C., Palazzo, S., He, J., Beauxis-Aussalet, E., Lin, S.I., Chou, H.M., Nadarajan, G., Chen-Burger, Y.H., van Ossensbruggen, J., Giordano, D., Hardman, L., Lin, F.P., Fisher, R.B., 2012. Long-term underwater camera surveillance for monitoring and analysis of fish populations. *Workshop on Visual observation and Analysis of Animal and Insect Behavior (VAIB)*, in conjunction with ICPR 2012 , 2–5 URL: <http://homepages.inf.ed.ac.uk/rbf/VAIB12PAPERS/boom.pdf>.
- [24] Botta, M., Giordana, A., Saitta, L., 1993. Learning fuzzy concept definitions. *1993 IEEE International Conference on Fuzzy Systems* , 18–22 doi:10.1109/fuzzy.1993.327470.
- [25] Bozcan, I., Kayacan, E., 2020. AU-AIR: A Multi-modal Unmanned Aerial Vehicle Dataset for Low Altitude Traffic Surveillance URL: <http://arxiv.org/abs/2001.11737>.
- [26] Braun, M., Krebs, S., Flohr, F., Gavrila, D.M., 2018. The EuroCity Persons Dataset: A Novel Benchmark for Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 1844–1861. URL: <http://arxiv.org/abs/1805.07193> <http://dx.doi.org/10.1109/TPAMI.2019.2897684>, doi:10.1109/TPAMI.2019.2897684.
- [27] Brostow, G., Shotton, J., Fauqueur, J., Cipolla, R., 2008a. Segmentation and Recognition using SfM Point Clouds. *Eccv* , 1–15.
- [28] Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R., 2008b. Segmentation and Recognition Using Structure from Motion Point Clouds , 44–57.
- [29] Brox, T., Malik, J., 2010. Object Segmentation by Long Term Analysis of Point Trajectories, in: Daniilidis, K., Maragos, P., Paragios, N. (Eds.), *Computer Vision – ECCV 2010*, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 282–295.
- [30] Caelles, S., Pont-Tuset, J., Perazzi, F., Montes, A., Maninis, K.K., Van Gool, L., 2019. The 2019 DAVIS Challenge on VOS: Unsupervised Multi-Object Segmentation , 1–4 URL: <http://arxiv.org/abs/1905.00737>.
- [31] Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O., 2019. nuScenes: A multimodal dataset for autonomous driving .
- [32] Caesar, H., Uijlings, J., Ferrari, V., 2018. COCO-Stuff Thing and Stuff Classes in Context - Caesar, Uijlings, Ferrari -

- 2016.pdf , 1209–1218 URL: http://openaccess.thecvf.com/content/{_}cvpr/{_}2018/html/Caesar{_}COCO-Stuff{_}Thing{_}and{_}CVPR{_}2018{_}paper.html.
- [33] Cai, Y., Osman, S., Sharma, M., Landis, M., Li, S., 2015. Multi-Modality Vertebra Recognition in Arbitrary Views Using 3D Deformable Hierarchical Model. *IEEE Transactions on Medical Imaging* 34, 1676–1693. doi:10.1109/TMI.2015.2392054.
 - [34] Caicedo, J.C., Goodman, A., Karhohs, K.W., Cimini, B.A., Ackerman, J., Haghighi, M., Heng, C., Becker, T., Doan, M., McQuin, C., Rohban, M., Singh, S., Carpenter, A.E., 2019. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nature Methods* 16, 1247–1253. URL: <https://doi.org/10.1038/s41592-019-0612-7>, doi:10.1038/s41592-019-0612-7.
 - [35] Canfield, Kittler, H., Codella, N., Celebi, M.E., Dana, K., Halpern, A., Helba, B., Tschandl, P., . ISIC 2018. URL: <https://challenge2018.isic-archive.com/>.
 - [36] Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A., 2018. VGGFace2: A dataset for recognising faces across pose and age, in: *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018, Institute of Electrical and Electronics Engineers Inc.* pp. 67–74. doi:10.1109/FG.2018.00020.
 - [37] Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., Hays, J., 2019. Argoverse: 3D tracking and forecasting with rich maps. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June*, 8740–8749. doi:10.1109/CVPR.2019.00895.
 - [38] Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Return of the devil in the details: Delving deep into convolutional nets. *BMVC 2014 - Proceedings of the British Machine Vision Conference 2014* , 1–11 doi:10.5244/c.28.6.
 - [39] Che, Z., Li, G., Li, T., Jiang, B., Shi, X., Zhang, X., Lu, Y., Wu, G., Liu, Y., Ye, J., 2019. D²-City: A Large-Scale Dashcam Video Dataset of Diverse Traffic Scenarios URL: <http://arxiv.org/abs/1904.01975>.
 - [40] Chellapilla, K., Puri, S., Simard, P., 2006. High Performance Convolutional Neural Networks for Document Processing, in: Lorette, G. (Ed.), *Tenth International Workshop on Frontiers in Handwriting Recognition*, Suvisoft, La Baule (France).
 - [41] Chen, B.C., Chen, C.S., Hsu, W., 2014a. Cross-Age Reference Coding for Age-Invariant Face Recognition and Retrieval , 768–783.
 - [42] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 834–848.
 - [43] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017b. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 834–848.
 - [44] Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017c. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* .
 - [45] Chen, X., Girshick, R., He, K., Dollár, P., 2019. TensorMask: A Foundation for Dense Object Segmentation .
 - [46] Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A., 2014b. Detect what you can: Detecting and representing objects using holistic models and body parts, in: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, USA. p. 1979–1986. URL: <https://doi.org/10.1109/CVPR.2014.254>, doi:10.1109/CVPR.2014.254.
 - [47] Cheng, G., Han, J., 2016. A survey on object detection in optical remote sensing images. doi:10.1016/j.isprsjprs.2016.03.014.
 - [48] Cheng, G., Han, J., Lu, X., 2017. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE* 105, 1865–1883. URL: <http://arxiv.org/abs/1703.00121> <http://dx.doi.org/10.1109/JPROC.2017.2675998>, doi:10.1109/JPROC.2017.2675998.
 - [49] Cheng, M.M., Mitra, N., Huang, X., Hu, S.M., 2013. Salientshape: Group saliency in image collections. *The Visual Computer* 30, 1–10. doi:10.1007/s00371-013-0867-4.
 - [50] Choi, Y., Kim, N., Hwang, S., Park, K., Yoon, J.S., An, K., Kweon, I.S., 2018. KAIST Multi-Spectral Day/Night Data Set for Autonomous and Assisted Driving. *IEEE Transactions on Intelligent Transportation Systems* 19, 934–948. doi:10.1109/TITS.2018.2791533.
 - [51] Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua*, 1800–1807. doi:10.1109/CVPR.2017.195.
 - [52] Christie, G., Fendley, N., Wilson, J., Mukherjee, R., 2017. Functional Map of the World. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* , 6172–6180 URL: <http://arxiv.org/abs/1711.07846>.
 - [53] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., Halpern, A., 2019. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC) , 1–12.
 - [54] Coifman, B., Li, L., 2017. A critical evaluation of the Next Generation Simulation (NGSIM) vehicle trajectory dataset. *Transportation Research Part B: Methodological* 105, 362–377. doi:10.1016/j.trb.2017.09.018.
 - [55] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-Decem*, 3213–3223. doi:10.1109/CVPR.2016.350.
 - [56] Cuzzolin, F., Bawa, V.S., Skarga-Bandurova, I., Singh, G., 2020a. SARAS-ESAD 2020.
 - [57] Cuzzolin, F., Bawa, V.S., Skarga-Bandurova, I., Singh, G., 2020b. SARAS-ESAD Dataset. URL: <https://sar-as-esad.grand-challenge.org/Dataset/>.
 - [58] Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. *Proceedings - 2005 IEEE Computer*

- Society Conference on Computer Vision and Pattern Recognition, CVPR 2005 I, 886–893. doi:10.1109/CVPR.2005.177.
- [59] David, O., Bryan, H., Amirata, G., Matt P., L., Euan A., A., David H., L., James Y., Z., 2019. EchoNet-Dynamic Dataset.
- [60] Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R., Works, C., . DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images. Technical Report.
- [61] Deng, J., Dong, W., Socher, R., Li, L.J., Kai Li, Li Fei-Fei, 2010. ImageNet: A large-scale hierarchical image database , 248–255doi:10.1109/cvpr.2009.5206848.
- [62] DiDi, 2019. D2-City Detection Domain Adaptation Challenge.
- [63] Djavadifar, A., 2020. Automatic detection of geometrical anomalies in composites manufacturing : a deep learning-based computer vision approach. Ph.D. thesis.
- [64] Dollar, P., Wojek, C., Schiele, B., Perona, P., 2010. Pedestrian detection: A benchmark, Institute of Electrical and Electronics Engineers (IEEE). pp. 304–311. doi:10.1109/cvpr.2009.5206631.
- [65] ELCAP, 2003. ELCAP Public Lung Image Database. URL: <http://www.via.cornell.edu/lungdb.html>.
- [66] Enzweiler, M., Gavrilu, D.M., 2009. Monocular pedestrian detection: Survey and experiments, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 2179–2195. doi:10.1109/TPAMI.2008.260.
- [67] Etten, A.V., Lindenbaum, D., Bacastow, T., . SpaceNet: A Remote Sensing Dataset and Challenge Series. Technical Report.
- [68] Everingham, M., Eslami, S.M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2014. The Pascal Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision 111, 98–136. doi:10.1007/s11263-014-0733-5.
- [69] Everingham, M., Sivic, J., Zisserman, A., 2006. "Hello! My name is... Buffy" - Automatic naming of characters in TV video. BMVC 2006 - Proceedings of the British Machine Vision Conference 2006 , 899–908.
- [70] Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (VOC) challenge. International Journal of Computer Vision 88, 303–338. doi:10.1007/s11263-009-0275-4.
- [71] Fan, D.p., Guolei, G.p.J., Cheng, S.M.m., Shen, J., Shao, L., 2020a. Camouflaged Object Detection .
- [72] Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L., 2020b. Camouflaged object detection , 2774–2784doi:10.1109/CVPR42600.2020.00285.
- [73] Fan, D.P., Liu, J.J., Gao, S., Hou, Q., Borji, A., Cheng, M.M., 2018. Salient objects in clutter: Bringing salient object detection to the foreground. European Conference on Computer Vision (ECCV) .
- [74] Fan, Q., Zhong, F., Lischinski, D., Cohen-Or, D., Chen, B., 2015. JumpCut: Non-Successive Mask Transfer and Interpolation for Video Cutout. ACM Trans. Graph. 34. URL: <https://doi.org/10.1145/2816795.2818105>, doi:10.1145/2816795.2818105.
- [75] Fei- Fei, L., Fergus, R., Perona, P., 2004. Learning Generative Visual Models from Few Training Examples :. Conference on Computer Vision and Pattern Recognition Workshop (CVPR 2004) 00, 178. URL: <http://dx.doi.org/10.1109/CVPR.2004.109>, doi:10.1109/CVPR.2004.109.
- [76] Fellbaum, C., 1998. WordNet: an Electronic Lexical Database. Bradford Books.
- [77] Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1627–1645. doi:10.1109/TPAMI.2009.167.
- [78] Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Gläser, C., Timm, F., Wiesbeck, W., Dietmayer, K., 2021. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems 22, 1341–1360. doi:10.1109/TITS.2020.2972974.
- [79] Flanders, A.E., Prevedello, L.M., Shih, G., Halabi, S.S., Kalpathy-Cramer, J., Ball, R., Mongan, J.T., Stein, A., Kitamura, f.C., Lungren, Mattew, P., Choudhary, G., Cala.lesley, Coelho, L., Mogensen, M., Moron, F., Miller, E., Ikuta, I., Zohrabian, V., McDonnell, O., Lincoln, C., Shah, L., Joyner, D., Agarwal, A., Lee, R.K., Nath, J., . Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. URL: <https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection>, doi:<https://doi.org/10.1148/ryai.2020190211>.
- [80] Gan, D., Lin, G., Wu, H., Peng, J., Zhang, Y., Liao, B., Huang, Y., Zheng, Q., Zhang, N., 2017. Research and development of power grid dispatching operation control system based on transmission section control. Dianli Xitong Baohu yu Kongzhi/Power System Protection and Control 45, 117–124. doi:10.7667/PSPC161855.
- [81] Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Garcia-Rodriguez, J., 2017. A review on deep learning techniques applied to semantic segmentation .
- [82] Ge, Y., Zhang, R., Wang, X., Tang, X., Luo, P., 2019. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 5332–5340. doi:10.1109/CVPR.2019.00548.
- [83] Geiger, A., Lenz, P., Stiller, C., Urtasun, R., a. The KITTI 2D Object Evaluation Benchmark.
- [84] Geiger, A., Lenz, P., Stiller, C., Urtasun, R., b. The KITTI 3D Object Evaluation Benchmark.
- [85] Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: The KITTI dataset. International Journal of Robotics Research 32, 1231–1237. doi:10.1177/0278364913491297.
- [86] Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the KITTI vision benchmark suite. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition , 3354–3361doi:10.1109/CVPR.2012.6248074.
- [87] Girshick, R., 2015. Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, pp. 1440–1448.
- [88] Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580–587.

- [89] Goëau, H., Bonnet, P., Joly, A., 2019. Overview of LifeCLEF Plant identification task 2019: Diving into data deficient tropical countries, in: CEUR Workshop Proceedings, pp. 9–12.
- [90] Goldbaum, M., 1975. STARE Database.
- [91] Gould, S., Fulton, R., Koller, D., 2009. Decomposing a scene into geometric and semantically consistent regions. Proceedings of the IEEE International Conference on Computer Vision , 1–8doi:10.1109/ICCV.2009.5459211.
- [92] Griffin, Greg, 2007. Caltech-256 Object Category Dataset , 300.
- [93] Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J., 2016. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition .
- [94] Gupta, A., Dollar, P., Girshick, R., 2019. Lvis: A dataset for large vocabulary instance segmentation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June, 5351–5359. doi:10.1109/CVPR.2019.00550.
- [95] Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J., 2011. Semantic Contours from Inverse Detectors * - Hariharan et al.pdf. International Conference on Computer Vision , 8URL: <http://home.bharathh.info/pubs/pdfs/BharathICCV2011.pdf>.
- [96] He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969.
- [97] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society. pp. 770–778. URL: <http://image-net.org/challenges/LSVRC/2015/>, doi:10.1109/CVPR.2016.90.
- [98] Heath, M., Bowyer, K., Kopans, D., Morre, R., Kegelmeyer, W. Philip Chang, K., Munishkumaran, S., . Current Status of the Digital Database for Screening Mammography. Digital Mammography , 457–460doi:https://doi.org/10.1007/978-94-011-5318-8_75.
- [99] Heath, M., Bowyer, K., Kopans, D., Morre, R., Kegelmeyer, W. Philip Chang, K., Munishkumaran, S., 2001. THE DIGITAL DATABASE FOR SCREENING MAMMOGRAPHY. Medical Physics Publishing.
- [100] Heitz, G., Koller, D., . Learning Spatial Context: Using Stuff to Find Things. Technical Report.
- [101] Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., Dean, J., Tradewell, M., Shah, A., Tejapaul, R., Edgerton, Z., Peterson, M., Raza, S., Regmi, S., Papanikolopoulos, N., Weight, C., 2019. The KiTS19 Challenge Data: 300 Kidney Tumor Cases with Clinical Context, CT Semantic Segmentations, and Surgical Outcomes , 1–14.
- [102] Hinton, G.E., Osindero, S., Teh, Y.W., 2006. A fast learning algorithm for deep belief nets. Neural computation 18, 1527–1554.
- [103] Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. science 313, 504–507.
- [104] Hong-Wei, N., Stefan, W., 2014. A DATA-DRIVEN APPROACH TO CLEANING LARGE FACE DATASETS. International Conference on Image Processing(ICIP) , 343–347.
- [105] Horn, G.V., Aodha, O.M., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S., 2018. The iNaturalist Species Classification and Detection Dataset, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 8769–8778. doi:10.1109/CVPR.2018.00914.
- [106] Horn, G.V., Branson, S., Farrell, R., Barry, J., Tech, C., . Building a bird recognition app and large scale dataset with citizen scientists : The fine print in fine-grained dataset collection .
- [107] Hosseini, M.S., Chan, L., Tse, G., Tang, M., Deng, J., Norouzi, S., Rowsell, C., Plataniotis, K.N., Damaskinos, S., 2019. Atlas of digital pathology: A generalized hierarchical histological tissue type-annotated database for deep learning, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach. pp. 11739–11748. doi:10.1109/CVPR.2019.01202.
- [108] Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E., Learned-Miller, E., 2008. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report. URL: <https://hal.inria.fr/inria-00321923>.
- [109] Huang, J., Feris, R., Chen, Q., Yan, S., 2015. Cross-domain image retrieval with a dual attribute-aware ranking network, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1062–1070. doi:10.1109/ICCV.2015.127.
- [110] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y., 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Proceedings of the AAAI Conference on Artificial Intelligence 33, 590–597. doi:10.1609/aaai.v33i01.3301590.
- [111] Jacobs, C., Setio, A.A.A., Traverso, A., Ginneken, B.V., 2016. LUNA 2016.
- [112] Jain, S., Grauman, K., 2014. Supervoxel-Consistent Foreground Propagation in Video, pp. 656–671. doi:10.1007/978-3-319-10593-2_43.
- [113] Jesorsky, O., Kirchberg, K.J., Frischholz, R.W., 2001. Robust face detection using the Hausdorff distance. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2091, 90–95. doi:10.1007/3-540-45344-x_14.
- [114] Jonathon Phillips, P., Moon, H., Rizvi, S.A., Rauss, P.J., 2000. The FERET evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 1090–1104. doi:10.1109/34.879790.
- [115] Kaggle, 2018. CVPR 2018 WAD Video Segmentation Challenge. doi:<https://www.kaggle.com/c/cvpr-2018-autonomous-driving>.
- [116] Kaggle.com, 2017. Dstl satellite imagery feature detection. URL: <https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection>.
- [117] Kärkkäinen, K., Joo UCLA, J., . FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. Technical

- Report. URL: <https://github.com/joojs/fairface{\%}7D>.
- [118] Kauppi, T., Kalesnykiene, V., Kamarainen, J.K., Lensu, L., Sorri, I., Raninen, A., Voutilainen, R., Pietilä, J., Kälviäinen, H., Uusitalo, H., 2007. The DIARETDB1 diabetic retinopathy database and evaluation protocol. *BMVC 2007 - Proceedings of the British Machine Vision Conference 2007* , 1–18doi:10.5244/C.21.15.
 - [119] Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E., 2016. The MegaFace benchmark: 1 million faces for recognition at scale. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-Decem*, 4873–4882. doi:10.1109/CVPR.2016.527.
 - [120] Kesten, R., Usman, M., Houston, J., Pandya, T., Nadhamuni, K., Ferreira, A., Yuan, M., Low, B., Jain, A., Ondruska, P., Omari, S., Shah, S., Kulkarni, A., Kazakova, A., Tao, C., Platinsky, L., Jiang, W., Shet, V., 2019. Lyft Level 5 AV Dataset. URL: <https://level5.lyft.com/dataset/>.
 - [121] Khan, M.H., McDonagh, J., Khan, S., Shahabuddin, M., Arora, A., Khan, F.S., Shao, L., Tzimiropoulos, G., 2019. AnimalWeb: A Large-Scale Hierarchical Dataset of Annotated Animal Faces , 1–15URL: <http://arxiv.org/abs/1909.04951>.
 - [122] Khosla, A., Jayadevaprakash, N., Yao, B., Fei-Fei, L., 2011. Novel dataset for fine-grained image categorization. *Proc. IEEE Conf. Comput. Vision and Pattern Recognition* .
 - [123] Kiapour, M.H., Han, X., Lazebnik, S., Berg, A.C., Berg, T.L., 2015. Where to buy it: Matching street clothing photos in online shops, in: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3343–3351. doi:10.1109/ICCV.2015.382.
 - [124] Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Burge, M., Jain, A.K., 2015. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 07-12-June, 1931–1939*. doi:10.1109/CVPR.2015.7298803.
 - [125] Krajewski, R., Bock, J., Kloeker, L., Eckstein, L., 2018. The highD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC 2018-November, 2118–2125*. URL: <http://arxiv.org/abs/1810.05642>.
 - [126] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L., 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123, 32–73. doi:10.1007/s11263-016-0981-7.
 - [127] Krizhevsky, A., 2012. Learning Multiple Layers of Features from Tiny Images. University of Toronto .
 - [128] Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* , 1097–1105URL: <http://arxiv.org/abs/1102.0183>.
 - [129] Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K., 2009. Attribute and simile classifiers for face verification. *Proceedings of the IEEE International Conference on Computer Vision* , 365–372doi:10.1109/ICCV.2009.5459250.
 - [130] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallocci, M., Kolesnikov, A., Duerig, T., Ferrari, V., 2018. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale , 1–20URL: <http://arxiv.org/abs/1811.00982>.
 - [131] Lam, D., Kuzma, R., McGee, K., Dooley, S., Laielli, M., Klaric, M., Bulatov, Y., McCord, B., 2018. xView: Objects in Context in Overhead Imagery URL: <http://arxiv.org/abs/1802.07856>.
 - [132] Lambert, Z., Petitjean, C., Dubray, B., Ruan, S., 2019. SegTHOR: Segmentation of Thoracic Organs at Risk in CT images , 1–16.
 - [133] LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A., Raichle, Marcus, E., Carlos, C., Marcus, D., 2019. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. *Journal of Chemical Information and Modeling* 53, 1689–1699. doi:10.1017/CB09781107415324.004.
 - [134] Lateef, F., Ruichek, Y., 2019. Survey on semantic segmentation using deep learning techniques. *Neurocomputing* 338, 321–348. URL: <https://www.sciencedirect.com/science/article/pii/S092523121930181X>, doi:<https://doi.org/10.1016/j.neucom.2019.02.003>.
 - [135] Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2*, 2169–2178. doi:10.1109/CVPR.2006.68.
 - [136] Le, T.N., Nguyen, T., Nie, Z., Tran, M.T., Sugimoto, A., 2019. Anabranh network for camouflaged object segmentation. *Computer Vision and Image Understanding* 184. doi:10.1016/j.cviu.2019.04.006.
 - [137] Lecun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D., 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation* 1, 541–551.
 - [138] Lecun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D., 1990. Handwritten digit recognition with a back-propagation network, in: *Advances in neural information processing systems*, pp. 396–404.
 - [139] Lecun, Y., Bottou, L., Bengio, Y., Ha, P., 1998. LeNet. *Proceedings of the IEEE* , 1–46doi:10.1109/5.726791.
 - [140] Lecun, Y., Others, 1997. Handwritten Digit Recognition with a Back-Propagation Network. *Neural Information Processing Systems 2*.
 - [141] LERA, 2018. LERA- Lower Extremity RAdiographs. URL: <https://aimi.stanford.edu/lera-lower-extremity-radiographs-2>.
 - [142] Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M., 2013. Video Segmentation by Tracking Many Figure-Ground Segments, in: *2013 IEEE International Conference on Computer Vision*, pp. 2192–2199. doi:10.1109/ICCV.2013.273.
 - [143] Li, G., Yu, Y., 2015. Visual saliency based on multiscale deep features doi:10.1109/CVPR.2015.7299184.
 - [144] Li, H., Chen, M., 2020. Automatic Structure Segmentation for Radiotherapy Planning Challenge 2020. doi:10.5281/

- zenodo.3718885.
- [145] Li, H., Zhou, J., Deng, J., Chen, M., SenseTime, YINO, Zhejiang Cancer Hospital, 2019. StructSeg 2019.
 - [146] Li, M., Zang, S., Zhang, B., Li, S., Wu, C., 2014a. A review of remote sensing image classification techniques: the role of spatio-contextual information. *European Journal of Remote Sensing* 47, 389–411. URL: <https://doi.org/10.5721/EuJRS20144723>, doi:10.5721/EuJRS20144723.
 - [147] Li, S., Wang, 2019. AASCE. URL: <https://aasce19.grand-challenge.org/>.
 - [148] Li, X., Yang, F., Cheng, H., Chen, J., Guo, Y., Chen, L., 2017. Multi-scale cascade network for salient object detection , 439–447doi:10.1145/3123266.3123290.
 - [149] Li, X., Yang, F., Cheng, H., Liu, W., Shen, D., 2018. Contour knowledge transfer for salient object detection: 15th european conference, munich, germany, september 8-14, 2018, proceedings, part xv , 370–385doi:10.1007/978-3-030-01267-0_22.
 - [150] Li, Y., Hou, X., Koch, C., Rehg, J., Yuille, A., 2014b. The secrets of salient object segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* doi:10.1109/CVPR.2014.43.
 - [151] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2016. Feature pyramid networks for object detection .
 - [152] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8693 LNCS, 740–755. doi:10.1007/978-3-319-10602-1_48.
 - [153] Liu, C., Yuen, J., Torralba, A., 2015a. Nonparametric scene parsing via label transfer. *Dense Image Correspondences for Computer Vision* 33, 207–236. doi:10.1007/978-3-319-23048-1_10.
 - [154] Liu, K., Mattyus, G., 2015. Fast Multiclass Vehicle Detection on Aerial Images. *IEEE Geoscience and Remote Sensing Letters* 12, 1938–1942. doi:10.1109/LGRS.2015.2439517.
 - [155] Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M., 2020a. Deep Learning for Generic Object Detection: A Survey. *International Journal of Computer Vision* 128, 261–318. URL: <https://doi.org/10.1007/s11263-019-01247-4>, doi:10.1007/s11263-019-01247-4.
 - [156] Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikinen, M., 2020b. Deep learning for generic object detection: A survey. *International Journal of Computer Vision* 128, 261–318. URL: <https://doi.org/10.1007/s11263-019-01247-4>, doi:10.1007/s11263-019-01247-4.
 - [157] Liu, T., Sun, J., Zheng, N.N., Tang, X., Shum, H.Y., 2007. Learning to detect a salient object , 1–8doi:10.1109/CVPR.2007.383047.
 - [158] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2015b. SSD: Single Shot MultiBox Detector doi:10.1007/978-3-319-46448-0_2.
 - [159] Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X., 2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations, in: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1096–1104. doi:10.1109/CVPR.2016.124.
 - [160] Liu, Z., Luo, P., Wang, X., Tang, X., 2015c. Deep learning face attributes in the wild. *Proceedings of the IEEE International Conference on Computer Vision 2015 Inter*, 3730–3738. doi:10.1109/ICCV.2015.425.
 - [161] Lowe, D.G., 1999. Object recognition from local scale-invariant features, in: *Proceedings of the IEEE International Conference on Computer Vision, IEEE*, pp. 1150–1157. doi:10.1109/iccv.1999.790410.
 - [162] Lyft, 2019. Lyft 3D Object Detection for Autonomous Vehicles. URL: <https://www.kaggle.com/c/3d-object-detection-for-autonomous-vehicles>.
 - [163] Maddern, W., Pascoe, G., Linegar, C., Newman, P., . 1 Year , 1000km : The Oxford RobotCar Dataset 3.
 - [164] Maier, O., 2015. SMIR Database URL: <https://www.smir.ch>.
 - [165] Martin, D., Fowlkes, C., Tal, D., Malik, J., 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *Proceedings of the IEEE International Conference on Computer Vision* 2, 416–423. doi:10.1109/ICCV.2001.937655.
 - [166] Martinez, A.M., 1998. The AR face database. *CVC Technical Report24* .
 - [167] Masi, I., Wu, Y., Hassner, T., Natarajan, P., 2019. Deep Face Recognition: A Survey. *Proceedings - 31st Conference on Graphics, Patterns and Images, SIBGRAPI 2018* , 471–478doi:10.1109/SIBGRAPI.2018.00067.
 - [168] Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., Grother, P., 2018. IARPA janus benchmark-C: Face dataset and protocol. *Proceedings - 2018 International Conference on Biometrics, ICB 2018* , 158–165doi:10.1109/ICB2018.2018.00033.
 - [169] Merler, M., Ratha, N., Feris, R.S., Smith, J.R., 2019. Diversity in Faces , 1–29URL: <http://arxiv.org/abs/1901.10436>.
 - [170] Meyer, M., Kusch, G., 2019. Automotive radar dataset for deep learning based 3D object detection. *EuRAD 2019 - 2019 16th European Radar Conference* , 129–132.
 - [171] Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A., 2014. The role of context for object detection and semantic segmentation in the wild. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* , 891–898doi:10.1109/CVPR.2014.119.
 - [172] Mundhenk, T.N., Konjevod, G., Sakla, W.A., Boakye, K., 2016. A Large Contextual Dataset for Classification, Detection and Counting of Cars with Deep Learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9907 LNCS, 785–800. URL: <http://arxiv.org/abs/1609.04453>.
 - [173] National Library of Medicine, 2006. MedPix. URL: <https://medpix.nlm.nih.gov/home>.
 - [174] Nech, A., Kemelmacher-Shlizerman, I., Allen, P.G., . Level Playing Field for Million Scale Face Recognition. *Technical Report*.

- [175] Nene, S., Nayar, S., Murase, H., 1996a. Columbia Object Image Library (COIL-100). Technical Report 95, 223–303. URL: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.5914>.
- [176] Nene, S., Nayar, S., Murase, H., 1996b. Columbia Object Image Library (COIL-20). Technical Report 95, 223–303. URL: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.5914>.
- [177] Neuhold, G., Ollmann, T., Bulo, S.R., Kotschieder, P., 2017. The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. Proceedings of the IEEE International Conference on Computer Vision 2017-Octob, 5000–5009. doi:10.1109/ICCV.2017.534.
- [178] Neumann, L., Karg, M., Zhang, S., Scharfenberger, C., Piegert, E., Mistr, S., Prokofyeva, O., Thiel, R., Vedaldi, A., Zisserman, A., Schiele, B., 2019. NightOwls: A Pedestrians at Night Dataset, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag. pp. 691–705. URL: <http://www.nightowls-dataset.org/>, doi:10.1007/978-3-030-20887-5_43.
- [179] Nilsback, M.E., Zisserman, A., 2008. Automated flower classification over a large number of classes. Proceedings - 6th Indian Conference on Computer Vision, Graphics and Image Processing, ICVGIP 2008 , 722–729doi:10.1109/ICVGIP.2008.47.
- [180] Ochs, P., Malik, J., Brox, T., 2014. Segmentation of Moving Objects by Long Term Video Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 36, 1187–1200. doi:10.1109/TPAMI.2013.242.
- [181] Odir, 2019. ODIR-5K. URL: <http://www.kaggle.com/andrewmvd/ocular-disease-recognition-odir5k>.
- [182] Orlando, J.I., Fu, H., Breda, J.B., van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P., Kim, J., Lee, J., Lee, J., Li, X., Liu, P., Lu, S., Murugesan, B., Naranjo, V., Phay, S.S.R., Shankaranarayana, S.M., Sikka, A., Son, J., van den Hengel, A., Wang, S., Wu, J., Wu, Z., Xu, G., Xu, Y., Yin, P., Li, F., Zhang, X., Xu, Y., Bogunovic, H., 2019. REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. CoRR abs/1910.03667.
- [183] Osuna, E., Freund, R., Girosi, F., 1997. Training support vector machines: An application to face detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition , 130–136doi:10.1109/cvpr.1997.609310.
- [184] Papageorgiou, C., Poggio, T., 2000. Trainable system for object detection. International Journal of Computer Vision 38, 15–33. doi:10.1023/A:1008162616689.
- [185] Parkhi, O.M., Vedaldi, A., Zisserman, A., 2015. Deep Face Recognition , 41.1–41.12doi:10.5244/c.29.41.
- [186] Patil, A., Malla, S., Gang, H., Chen, Y.T., 2019. The H3D dataset for full-surround 3D multi-object detection and tracking in crowded urban scenes. Proceedings - IEEE International Conference on Robotics and Automation 2019-May, 9552–9557. doi:10.1109/ICRA.2019.8793925.
- [187] Patterson, G., Hays, J., 2012. SUN attribute database: Discovering, annotating, and recognizing scene attributes. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition , 2751–2758doi:10.1109/CVPR.2012.6247998.
- [188] Pham, Q.H., Sevestre, P., Pahwa, R.S., Zhan, H., Pang, C.H., Chen, Y., Mustafa, A., Chandrasekhar, V., Lin, J., 2019. A*3D Dataset: Towards Autonomous Driving in Challenging Environments .
- [189] Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.J., 1998. The FERET database and evaluation procedure for face-recognition algorithms. Image and Vision Computing 16, 295–306. doi:10.1016/s0262-8856(97)00070-x.
- [190] Prest, A., Leistner, C., Civera, J., Schmid, C., Ferrari, V., 2012. Learning object class detectors from weakly annotated video. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition , 3282–3289doi:10.1109/CVPR.2012.6248065.
- [191] Quattoni, A., Torralba, A., 2010. Recognizing indoor scenes. 2009 IEEE Conference on Computer Vision and Pattern Recognition , 413–420doi:10.1109/cvpr.2009.5206537.
- [192] Radiological Society of North America, 2019. RSNA Intracranial Hemorrhage Detection.
- [193] Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R.L., Langlotz, C., Shpanskaya, K., Lungren, M.P., Ng, A.Y., 2017. MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs , 1–10.
- [194] Ranzato, M., Huang, F.J., Boureau, Y., LeCun, Y., 2007. Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. doi:10.1109/CVPR.2007.383157.
- [195] Rawat, W., Wang, Z., 2017. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. Neural Computation 29, 2352–2449. doi:10.1162/neco_a_00990.
- [196] Razakarivony, S., Jurie, F., 2016. Vehicle detection in aerial imagery: A small target detection benchmark. Journal of Visual Communication and Image Representation 34, 187–203. doi:10.1016/j.jvcir.2015.11.002.
- [197] Real, E., Shlens, J., Mazzocchi, S., Pan, X., Vanhoucke, V., 2017. YouTube-BoundingBoxes: A large high-precision human-annotated data set for object detection in video, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 7464–7473. doi:10.1109/CVPR.2017.789.
- [198] Redmon, J., Divvala, S., Girshick, R., Farhadi, A., . You Only Look Once: Unified, Real-Time Object Detection.
- [199] Redmon, J., Farhadi, A., 2016. YOLO9000: Better, Faster, Stronger .
- [200] Redmon, J., Farhadi, A., 2018. YOLOv3: An Incremental Improvement .
- [201] Reinertsen, I., Xiao, Y., Rivaz, H., Chabanas, M., 2019. CuRIOUS 2019. URL: <https://curious2019.grand-challenge.org/>.
- [202] Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, pp. 91–99.
- [203] Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in:

- International Conference on Medical image computing and computer-assisted intervention, pp. 234–241.
- [204] Rothe, R., Timofte, R., Van Gool, L., . Deep expectation of real and apparent age from a single image without facial landmarks Real age 20 years DEX age predic3on. Technical Report.
 - [205] Rothe, R., Timofte, R., Van Gool, L., 2018. Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks. *International Journal of Computer Vision* 126, 144–157. doi:10.1007/s11263-016-0940-3.
 - [206] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 211–252. doi:10.1007/s11263-015-0816-y.
 - [207] Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T., 2008. LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision* 77, 157–173. doi:10.1007/s11263-007-0090-8.
 - [208] Schroff, F., Philbin, J., . FaceNet: A Unified Embedding for Face Recognition and Clustering. Technical Report.
 - [209] Sensing, R., Sciences, S.I., Hackel, T., Savinov, N., Ladicky, L., Wegner, J.D., Schindler, K., Pollefeys, M., 2017. SEMANTIC3D . NET : A NEW LARGE-SCALE POINT CLOUD CLASSIFICATION IV, 6–9. doi:10.5194/isprs-annals-IV-1-W1-91-2017.
 - [210] Shafiee, M.J., Chywl, B., Li, F., Wong, A., 2017. Fast YOLO: A Fast You Only Look Once System for Real-time Embedded Object Detection in Video .
 - [211] Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J., Technology, M., 2019. Objects365: A Large-scale, High-quality Dataset for Object Detection. *Proc. IEEE International Conference on Computer Vision (ICCV)* , 8430–8439.
 - [212] Shotton, J., Winn, J., Rother, C., Criminisi, A., 2006. TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation, in: Leonardis, A., Bischof, H., Pinz, A. (Eds.), *Computer Vision – ECCV 2006*, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 1–15.
 - [213] Silberman, N., Hoiem, D., Kohli, P., Fergus, R., 2012. Indoor segmentation and support inference from RGBD images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7576 LNCS, 746–760. doi:10.1007/978-3-642-33715-4_54.
 - [214] Sim, T., Baker, S., Bsat, M., 2002. The CMU Pose, Illumination, and Expression (PIE) database. *Proceedings - 5th IEEE International Conference on Automatic Face Gesture Recognition, FGR 2002* , 53–58doi:10.1109/AFGR.2002.1004130.
 - [215] Sirinukunwattana, K., Raza, s.E.A., Tsang, Y., Snead, D.R., Cree, I.A., Rajpoot, N.M., 2016. Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. *IEEE Transactions on Medical Imaging* 35, 1196–1206. doi:10.1109/TMI.2016.2525803.
 - [216] Song, S., Lichtenberg, S.P., Xiao, J., 2015. SUN RGB-D: A RGB-D scene understanding benchmark suite. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 07-12-June*, 567–576. doi:10.1109/CVPR.2015.7298655.
 - [217] Sørensen, L., Shaker, S.B., De Bruijne, M., 2010. Quantitative analysis of pulmonary emphysema using local binary patterns. *IEEE Transactions on Medical Imaging* 29, 559–569. doi:10.1109/TMI.2009.2038575.
 - [218] Souza, R., Lucena, O., Garrafa, J., Gobbi, D., Saluzzi, M., Appenzeller, S., Rittner, L., Frayne, R., Lotofo, R., 2018. An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement. *NeuroImage* 170, 482–494. doi:https://doi.org/10.1016/j.neuroimage.2017.08.021.
 - [219] Staal, J., Abràmoff, M., Niemeijer, M., Viergever, M., Ginneken, B., 2013. Digital Retinal Image for Vessel Extraction (DRIVE) Database.
 - [220] Sun, M., Yuan, Y., Zhou, F., Ding, E., 2018. Multi-Attention Multi-Class Constraint for Fine-grained Image Recognition, in: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 834–850. doi:10.1007/978-3-030-01270-0_49.
 - [221] Sun, P., Kretschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D., 2019. Scalability in Perception for Autonomous Driving: Waymo Open Dataset .
 - [222] Sun, Y., Liang, D., Wang, X., Tang, X., 2015. DeepID3: Face Recognition with Very Deep Neural Networks URL: <http://arxiv.org/abs/1502.00873>.
 - [223] Sun, Y., Wang, X., Tang, X., . Deep Learning Face Representation by Joint Identification-Verification. Technical Report.
 - [224] Sun, Y., Wang, X., Tang, X., 2014. Deep learning face representation from predicting 10,000 classes, in: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society. pp. 1891–1898. doi:10.1109/CVPR.2014.244.
 - [225] Sung, K.k., 1996. Learning and Example Selection for Object and Pattern Detection. PhD thesis , 195doi:https://doi.org/10.1016/j.comnet.2014.12.002.
 - [226] Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., Packer, C., 2015. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific Data* 2, 1–14. doi:10.1038/sdata.2015.26.
 - [227] Taghanaki, S.A., Abhishek, K., Cohen, J.P., Cohen-Adad, J., Hamarneh, G., 2020. Deep semantic segmentation of natural and medical images: A review .
 - [228] Taigman, Y., Marc', M.Y., Ranzato, A., Wolf, L., . DeepFace: Closing the Gap to Human-Level Performance in Face Verification. Technical Report.
 - [229] Taskiran, M., Kahraman, N., Erdem, C.E., 2020. Face recognition: Past, present and future (a review). *Digital Signal Processing* 106, 102809. URL: <https://www.sciencedirect.com/science/article/pii/S1051200420301548>, doi:https://doi.org/10.1016/j.dsp.2020.102809.
 - [230] Thomee, B., Elizalde, B., Shamma, D.A., Ni, K., Friedland, G., Poland, D., Borth, D., Li, L.J., 2016. YFCC100M: The

- new data in multimedia research. *Communications of the ACM* 59, 64–73. doi:10.1145/2812802.
- [231] Tighe, J., Lazebnik, S., 2010. SuperParsing: Scalable Nonparametric Image Parsing with Superpixels, in: Daniilidis, K., Maragos, P., Paragios, N. (Eds.), *Computer Vision – ECCV 2010*, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 352–365.
 - [232] Tighe, J., Lazebnik, S., 2013. Superparsing: Scalable nonparametric image parsing with superpixels. *International Journal of Computer Vision* 101, 329–349. doi:10.1007/s11263-012-0574-z.
 - [233] Torralba, A., Fergus, R., Freeman, W.T., 2008. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 1958–1970. doi:10.1109/TPAMI.2008.128.
 - [234] Torralba, A., Murphy, K.P., Freeman, W.T., 2004. Sharing features: Efficient boosting procedures for multiclass object detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2. doi:10.1109/cvpr.2004.1315241.
 - [235] Tschandl, P., Rosendahl, C., Kittler, H., 2018. Data descriptor: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* 5, 1–9. doi:10.1038/sdata.2018.161.
 - [236] Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N., 2017. EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. *IEEE Transactions on Medical Imaging* 36, 86–97. doi:10.1109/TMI.2016.2593957.
 - [237] University of Minnesota, University of Melbourne, 2019. KiTS19 Challenge. URL: <https://kits19.grand-challenge.org/>.
 - [238] Van Brummelen, J., O'Brien, M., Gruyer, D., Najjaran, H., 2018. Autonomous vehicle perception: The technology of today and tomorrow. *Transportation Research Part C: Emerging Technologies* 89, 384–406. URL: <https://doi.org/10.1016/j.trc.2018.02.012>, doi:10.1016/j.trc.2018.02.012.
 - [239] Viola, P., Viola, P., Jones, M., 2001a. Rapid object detection using a boosted cascade of simple features. *ACCEPTED CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION* 2001 .
 - [240] Viola, P., Jones, M., Jones, M., 2001b. Robust Real-time Object Detection. *INTERNATIONAL JOURNAL OF COMPUTER VISION* .
 - [241] Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S., 2011. The Caltech-ucsd Birds-200-2011 Dataset .
 - [242] Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X., 2017a. Learning to detect salient objects with image-level supervision , 3796–3805doi:10.1109/CVPR.2017.404.
 - [243] Wang, P., Huang, X., Cheng, X., Zhou, D., Geng, Q., Yang, R., 2019. The ApolloScape Open Dataset for Autonomous Driving and its Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 1–1doi:10.1109/tpami.2019.2926463.
 - [244] Wang, S., Bai, M., Mattyus, G., Chu, H., Luo, W., Yang, B., Liang, J., Cheverie, J., Fidler, S., Urtasun, R., . TorontoCity: Seeing the World with a Million Eyes. Technical Report.
 - [245] Wang, S., Bai, M., Mattyus, G., Chu, H., Luo, W., Yang, B., Liang, J., Cheverie, J., Fidler, S., Urtasun, R., 2016. TorontoCity : Seeing the World with a Million Eyes .
 - [246] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017b. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua*, 3462–3471. doi:10.1109/CVPR.2017.369.
 - [247] Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J., Miller, T., Kalka, N., Jain, A.K., Duncan, J.A., Allen, K., Cheney, J., Grother, P., 2017. IARPA Janus Benchmark-B Face Dataset. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* 2017-July, 592–600. doi:10.1109/CVPRW.2017.87.
 - [248] Winship Cancer Institute, . Cancer Digital Slide Archive. URL: <https://cancer.digitalslidearchive.org/>.
 - [249] Wolf, L., Hassner, T., Maoz, I., 2011. Face recognition in unconstrained videos with matched background similarity. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* , 529–534doi:10.1109/CVPR.2011.5995566.
 - [250] Wrenninge, M., Unger, J., 2018. Synscapes: A Photorealistic Synthetic Dataset for Street Scene Parsing URL: <http://arxiv.org/abs/1810.08705>.
 - [251] Wu, H., Bailey, C., Rasoulinejad, P., Li, S., 2017. Automatic Landmark Estimation for Adolescent Idiopathic Scoliosis Assessment Using BoostNet, in: *Medical Image Computing and Computer Assisted Intervention - MICCAI*, pp. 127–135.
 - [252] Wu, X., Zhan, C., Lai, Y.K., Cheng, M.M., Yang, J., 2019. IP102: A large-scale benchmark dataset for insect pest recognition, in: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 8779–8788. doi:10.1109/CVPR.2019.00899.
 - [253] Xia, C., Li, J., Chen, X., Zheng, A., Zhang, Y., 2017a. What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors , 4399–4407doi:10.1109/CVPR.2017.468.
 - [254] Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L., 2017b. DOTA: A Large-scale Dataset for Object Detection in Aerial Images. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* , 3974–3983URL: <http://arxiv.org/abs/1711.10398>.
 - [255] Xia, G.S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., 2016. AID: A Benchmark Dataset for Performance Evaluation of Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing* 55, 3965–3981. URL: <http://arxiv.org/abs/1608.05167><http://dx.doi.org/10.1109/TGRS.2017.2685945>, doi:10.1109/TGRS.2017.2685945.
 - [256] Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A., 2010. SUN database: Large-scale scene recognition from abbey to zoo. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* , 3485–3492doi:10.1109/CVPR.2010.5539970.

- [257] Xiao, Y., Fortin, M., Unsgård, G., Rivaz, H., Reinertsen, I., 2017. REtroSpective Evaluation of Cerebral Tumors (RESECT): A clinical database of pre-operative MRI and intra-operative ultrasound in low-grade glioma surgeries: A. Medical Physics 44, 3875–3882. doi:10.1002/mp.12268.
- [258] Xu, D., Anguelov, D., Jain, A., 2018a. PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation. Technical Report.
- [259] Xu, G., Song, Z., Sun, Z., Ku, C., Yang, Z., Liu, C., Wang, S., Ma, J., Xu, W., 2019. CAMEL: A weakly supervised learning framework for histopathology image segmentation. Proceedings of the IEEE International Conference on Computer Vision 2019-Octob, 10681–10690. doi:10.1109/ICCV.2019.01078.
- [260] Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T., 2018b. YouTube-VOS: A Large-Scale Video Object Segmentation Benchmark , 1–10URL: <http://arxiv.org/abs/1809.03327>.
- [261] Yan, Q., Shi, J., Xu, L., Jia, J., 2014. Hierarchical saliency detection on extended cssd. IEEE Transactions on Pattern Analysis and Machine Intelligence 38. doi:10.1109/TPAMI.2015.2465960.
- [262] Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H., 2013. Saliency detection via graph-based manifold ranking. Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition , 3166–3173doi:10.1109/CVPR.2013.407.
- [263] Yao, J., Burns, J.E., Forsberg, D., Seitel, A., Rasouljan, A., Abolmaesumi, P., Hammernik, K., Urschler, M., Ibragimov, B., Korez, R., Vrtovec, T., Castro-Mateos, I., Pozo, J.M., Frangi, A.F., Summers, R.M., Li, S., 2016. A multi-center milestone study of clinical vertebral CT segmentation. Computerized Medical Imaging and Graphics 49, 16–28. doi:10.1016/j.compmedimag.2015.12.006.
- [264] Yi, D., Lei, Z., Liao, S., Li, S.Z., 2014. Learning Face Representation from Scratch URL: <http://arxiv.org/abs/1411.7923>.
- [265] Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., Darrell, T., 2018. BDD100K: A Diverse Driving Video Database with Scalable Annotation Tooling , 1–16.
- [266] Zhai, Q., Li, X., Yang, F., Chen, C., Cheng, H., Fan, D.P., 2021. Mutual graph learning for camouflaged object detection .
- [267] Zhan, W., Sun, L., Wang, D., Shi, H., Clausse, A., Naumann, M., Kummerle, J., Konigshof, H., Stiller, C., de La Fortelle, A., Tomizuka, M., 2019. INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps .
- [268] Zhang, J., Ma, S., Sameki, M., Sclaroff, S., Betke, M., Lin, Z., Shen, X., Price, B., Mech, R., 2015. Salient object subitizing , 4045–4054doi:10.1109/CVPR.2015.7299031.
- [269] Zhang, L., Zhang, J., Lin, Z., Lu, H., He, Y., 2019. Capsal: Leveraging captioning to boost semantics for salient object detection , 6017–6026doi:10.1109/CVPR.2019.00618.
- [270] Zhang, S., Benenson, R., Schiele, B., 2017. CityPersons: A Diverse Dataset for Pedestrian Detection. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-January, 4457–4465. URL: <http://arxiv.org/abs/1702.05693>.
- [271] Zhao, Z.Q., Zheng, P., Xu, S.T., Wu, X., 2019. Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems 30, 3212–3232. doi:10.1109/TNNLS.2018.2876865.
- [272] Zheng, S., Hadi Kiapour, M., Yang, F., Piramuthu, R., 2018. ModaNet: A large-scale street fashion dataset with polygon annotations. MM 2018 - Proceedings of the 2018 ACM Multimedia Conference , 1670–1678doi:10.1145/3240508.3240652.
- [273] Zhou, B., Lapedriza, A., Torralba, A., Oliva, A., 2017a. Places: An Image Database for Deep Scene Understanding. Journal of Vision 17, 296. doi:10.1167/17.10.296.
- [274] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A., 2017b. Scene parsing through ADE20K dataset. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua, 5122–5130. doi:10.1109/CVPR.2017.544.
- [275] Zhou, E., Cao, Z., Yin, Q., 2015. Naive-Deep Face Recognition: Touching the Limit of LFW Benchmark or Not? URL: <http://arxiv.org/abs/1501.04690>.
- [276] Zhou, E., Yin, Q., . Naive-Deep Face Recognition: Touching the Limit of LFW Benchmark or Not? Technical Report.
- [277] Zhu, H., Chen, X., Dai, W., Fu, K., Ye, Q., Jiao, J., 2015. Orientation robust object detection in aerial images using deep convolutional neural network, in: Proceedings - International Conference on Image Processing, ICIP, IEEE Computer Society. pp. 3735–3739. doi:10.1109/ICIP.2015.7351502.
- [278] Zou, X., Kong, X., Wong, W., Wang, C., Liu, Y., Cao, Y., 2019a. FashionAI: A Hierarchical Dataset for Fashion Understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops .
- [279] Zou, Z., Shi, Z., 2018. Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images. IEEE Transactions on Image Processing 27, 1100–1111. doi:10.1109/TIP.2017.2773199.
- [280] Zou, Z., Shi, Z., Guo, Y., Ye, J., 2019b. Object Detection in 20 Years: A Survey , 1–39URL: <http://arxiv.org/abs/1905.05055>.